The Dissertation Committee for Amir Shahmoradi
certifies that this is the approved version of the following dissertation:

Dissecting the relationship between protein structure and sequence evolution

Committee:

Swadesh M. Mahajan, Supervisor

Claus O. Wilke, Co-Supervisor

Raymond L. Orbach

Michael P. Marder

Vernita D. Gordon

William H. Press

Dissecting the relationship between protein structure and sequence evolution

by

Amir Shahmoradi, B.S.; M.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2015

To my father, a science lover and curious mind,

who showed me the path to science and the pleasure of finding things out.

To my mother who, during her short life,

showed me the key to success: hard work, diligence and love for my dreams.

Acknowledgments

Throughout my scientific life, I have been blessed to meet and work with some of the greatest contemporary minds and brightest scientists in the world. First and foremost, I thank my advisors Dr. Claus Wilke and Dr. Swadesh Mahajan. Having spent more than a decade within academia and higher education, I cannot not imagine any academic advisers who could be more supportive and caring about their students as they have been to me. No student could expect anything more.

In addition, I acknowledge Dr. William Press and Dr. Raymond Orbach whose insights into my research proved very helpful and enlightening. These insights have not only improved the research presented in this dissertation but also my research in the field of High Energy Astrophysics. Their rigor and enthusiasm in scientific discussions and teaching has been a source inspiration for me over the past years.

I thank Dr. Richard Fitzpatrick for his excellence in teaching and his role in my greater understanding of Quantum Mechanics and Electrodynamics. I also acknowledge my former adviser, Dr. Robert Nemiroff at Michigan Tech University, who taught me the basic elements of scientific research.

Several WilkeLab comrades, in particular Austin Meyer, Eleisha Jackson, Stephanie Spielman, and Dakota Derryberry were critical in my research

and training related to the fields of biology and biochemistry, in which I had no research experience prior to joining the WilkeLab.

Dissecting the relationship between protein structure and sequence evolution

Amir Shahmoradi, Ph.D.
The University of Texas at Austin, 2015

Supervisors: Swadesh M. Mahajan
Claus O. Wilke

What can protein structure tell us about protein evolutionary dynamics? Despite extensive variety in their native structures, from hyper-thermostable to intrinsically disordered, all proteins share a common feature: flexibility and dynamics at different levels of structure. In addition to spatial dynamics, proteins are also highly evolutionary dynamic polymers, exhibiting variability in their amino acid sequences on evolutionary timescales. Significant variations can be observed in the amino acid sequences of the divergent members of a single protein family, while their native conformations and biological functions remain almost conserved among all members of the family. These evolutionary variations can be due to a combination of point mutations, insertions, deletions or sometimes the rearrangement of domains in the protein sequence. In recent years, it has become increasingly evident that the dynamics of proteins

in space and time domains – corresponding to structural and evolutionary variations – mutually influence each other at the amino acid level. In particular, it is generally observed that the amino acids in the core of protein are more conserved than the amino acids on the surface. Some site-specific structural quantities have been already identified that are capable of explaining the general patterns of sequence variability in globular proteins. A prominent example is the amino acid exposure to solvent molecules – typically water – which surround proteins in vivo. Furthermore, some partial associations between the local flexibility, packing density and sequence variability can be also observed among globular proteins. There is however no consensus as to which set of structural characteristics play the dominant role in sequence evolution. The strength of sequence–structure correlations also appear to vary widely from one protein to another, with Spearman's correlation strength $\rho \in [0.1, 0.8]$.

Throughout a series of works summarized in the following chapters, first I explore the wide spectrum of structural determinants of sequence evolution, their interrelationships, and their role in the evolutionary dynamics of protein. I find that amino acid sites that are important for the overall stability of protein structure in general tend to be highly conserved. In other words, any amino acid substitution that results in a significant change of the potential energy landscape and thus the native conformation of protein, is disruptive and hence occurs less frequently on evolutionary timescale. I also find that long-range interactions among individual amino acids play a weak but non-negligible role in site-specific evolution of proteins and their inclusion

generally results in better predictions of sequence evolution from protein structure. Then, I present the results from a comprehensive search for the potential biophysical and structural determinants of protein evolution by studying > 200 structural and evolutionary characteristics of proteins in a dataset of viral and enzymatic proteins. I discuss the main protein properties responsible for the general patterns of protein evolution, and identify sequence divergence as the main determinant of the strengths of virtually all structure-evolution relationships, explaining $\sim 10 - 30\%$ of the observed variation in sequence-structure relations. In addition to sequence divergence, I identify several protein structural properties that are moderately but significantly coupled with the strength of sequence-structure relations. In particular, proteins with more homogeneous back-bone hydrogen bond energies, corresponding to proteins containing large fractions of helical secondary structures and low fraction of beta sheets tend to have the strongest sequence-structure relations.

Table of Contents

Acknowledgments **v**

Abstract **vii**

List of Tables **xiii**

List of Figures **xiv**

Chapter 1. Introduction **1**

Chapter 2. Site-Specific Structural and Evolutionary Characteristics of Proteins **6**

2.1 Site-Specific Measures of Sequence Variability 7

2.1.1 Codon models of evolutionary rates 10

2.1.2 Amino acid models of evolutionary rates 12

2.2 Site-Specific Sequence Variability as Measured from Protein Design 15

2.3 Site-Specific Stability Contribution to Protein Native Conformation 16

2.4 Site-Specific Solvent-Accessible Surface Area 17

2.5 Site-Specific Flexibility and Fluctuation Measures in Proteins . 18

2.5.1 Thermal Atomic Fluctuations as Proxy Measures of Amino Acid Flexibility in Proteins 21

2.5.2 Site-Specific Fluctuations from Protein Conformational Ensemble 22

2.6 Local Packing Density 26

2.7 Amino Acid Hydrogen Bond Strength 30

Chapter 3. Structural Determinants of Sequence Evolution in Viral Proteins: Buriedness, Packing, Flexibility, and Design **32**

3.1 Introduction 32

3.2 Materials and Methods 34

3.2.1 Sequence Preparation, Alignments, and the Calculation of Evolutionary Rates 34

3.2.2 Protein Crystal Structures 36

3.2.3 Molecular Dynamics Simulations 37

3.2.4 Measures of Buriedness, Packing Density, and Structural Flexibility 37

3.2.5 Sequence Entropy from Designed Proteins 39

3.3 Results 40

3.3.1 Dataset and Structural Variables Considered 40

3.3.2 Evaluating Structural Predictors of Sequence Evolution 43

3.3.3 MD Time-Averages vs. Crystal-Structure Snapshots . . 46

3.3.4 Sequence Entropy vs. Evolutionary-Rate Ratio ω 51

3.3.5 Multivariate Analysis of Structural Predictors 53

3.4 Discussion 57

Chapter 4. Structural Determinants of Sequence Evolution in Enzymatic proteins **64**

4.1 Introduction 64

4.2 Protein Dataset and Structure/Sequence Variability Measures 68

4.3 Voronoi Partitioning of Protein's Structure 71

4.3.1 Voronoi Cell Area and Volume as Proxy Measures of Local Packing Density and Flexibility in Proteins 73

4.4 Average Side Chain coordinates as the Best Representation of Protein 3D Structure 77

4.5 Discussion 83

4.5.1 Side-Chain vs. C_α B Factors in Representing Site-Specific Fluctuations 94

4.5.2 Long-Range Amino Acid Interactions Effects on Sequence Evolution 100

Chapter 5. Identifying the Structural and Evolutionary Modulators of the Strength of Sequence-Structure Relations 108

5.1 Introduction 108

5.2 Materials and Methods 110

5.2.1 Sequence Data, Alignments and Evolutionary Rates . . 110

5.2.2 Structural Properties 113

5.2.3 Eliminating Degeneracy in Structural Property Definitions 115

5.3 Results 116

5.3.1 Sequence Divergence as the Main Determinant of Sequence-Structure Relation 116

5.4 Discussion 121

Chapter 6. Conclusion 125

Bibliography 128

Vita 150

List of Tables

3.1 PDB structures considered in this study. 36

3.2 Availability of homologous crystal structures. Although most viral proteins have many PDB structures available, the sequence divergence among these structures is low. Therefore, when calculating RMSF from crystal structures, I considered only those proteins with at least five homologous structures at 5% pairwise sequence divergence (highlighted in bold). 42

3.3 Correlations between quantities obtained from MD trajectories and from crystal structures. For each quantity and each protein, I calculated the Spearman correlation ρ between the values obtained from MD time averages and the values obtained from viral protein crystal structures. Note that crystal structures for all nine proteins were used for RSA, CN, and WCN calculations, but only the six proteins for which I had sufficient crystal structure variability were used for CS RMSF. I then calculated the minimum, maximum, mean, and standard deviation of these correlations. 49

4.1 Best free parameters of different definitions of WCN (using four different weighting functions: power-law, exponential, Gaussian, and cutoff distance) that result in the strongest median Spearman's correlation (ρ) of WCN with four site-specific quantities (average side-chain B factor, evolutionary rates (r4sJC), sequence entropy, and $\Delta\Delta G$ rate) for the entire dataset of 209 proteins. The corresponding median correlation coefficients (ρ) are reported inside parenthesis next to each parameter value in the table. 103

List of Figures

2.1 An example four-taxon unrooted tree taken from [72] for illustration purposes. The external nodes (i.e., leaves) are labelled from 1 to 4 while internal nodes are 5 & 6. Branch lengths are denoted by t_i and the capital letters in parentheses on each node represent the one-letter abbreviations for the amino acids. 14

2.2 An illustration of methodology that is typically used for the calculation of Solvent Accessible Surface Areas of amino acids in individual sites in proteins. Depicted in this figure is the Glutamine molecule surrounded by solvent molecules (typically water) represented by the red spheres. For better illustration, the solvent molecules in front of the Glutamine have been removed. An approximate measure of solvent accessibility can be obtained by counting the number of spherically-shaped solvent molecules of radius $\sim 1.5\mathring{A}$ that can fit around an amino acids in a given site in protein. The solvent accessibility is therefore a discrete quantity by definition. (Illustration is courtesy of Austin G. Meyer, e.g., [118]) 19

2.3 A general positive trend between the Relative Solvent Accessibility of an amino acid in protein and its sequence evolutionary rates is normally seen in all proteins. Amino acids with RSA $\lesssim 0.2$ are considered to be buried deep in the core of protein, whereas sites with RSA$\gtrsim 0.2$ are considered to be part of the surface of the protein. The average curve shown in the plot was obtained by adjacent averaging over all sites in a dataset of 213 monomeric enzymes (c.f., Chapter 4). 20

2.4 A cartoon representation of Influenza Hemagglutinin protein *1RD8_AB*, illustrating the collective motions in secondary structures (i.e., α-helices and β-sheets). Color coding begins from the start (N-terminus) of the sequence in blue to the end (C-terminus) in red. The figure is a result of superposition of 12 conformational snapshots obtained every $100ps$ from Molecular Dynamics simulation of 1RD8_AB. The collective motion of amino acids in secondary structures, can potentially introduce strong biases in estimations of RMSF of amino acids in individual sites. 24

2.5 An illustration of the linear relationship between the standardized logarithm of local packing density as measured by the Weighted Contact Number (WCN) and the standardized logarithm of local flexibility of individual sites in proteins, as measured by the average side-chain B factor. The red curve is the result of adjacent averaging of 77150 amino acid sites in 209 enzymatic proteins (c.f., Chapter 4), every 3000 sites. The blue shaded area is 2D heat map of the scatter of 77150 sites about the average red curve. The dashed yellow line a symmetric Deming linear regression fit to the average curve, with a slope of $m \sim -0.7$. The observed deviation in the value of the slope from -1 is potentially the result of several contributing factors, most importantly, the systematic biases and errors in B factors as measures of local flexibility and the approximations made in obtaining the relationship between packing density and flexibility (Equations 2.20 & 2.21). 31

3.1 Spearman correlation of sequence entropy with measures of structural variability. Each symbol represents one correlation coefficient for one protein structure. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. The quantities $\mathrm{Var}(\psi)$, $\mathrm{Var}(\phi)$, $\mathrm{Var}(\chi_1)$, and MD RMSF were obtained as time-averages over 15ns of MD simulations. B factors were obtained from individual crystal structures. CS RMSF values were obtained from alignments of homologous crystal structures when available. Almost all structural measures of variability correlate weakly, but significantly, with sequence entropy. 45

3.2 Spearman correlation of sequence entropy with measures of buriedness, packing density, and structural flexibility, as well as with designed entropy. Each symbol represents one correlation coefficient for one protein structure. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. The quantities MD RSA, MD iWCN, MD $\mathrm{Var}(\chi_1)$, and MD RMSF were calculated as time-averages over 15ns of MD simulations. B factors were obtained from crystal structures, and designed entropy was obtained from protein design in Rosetta. Compared to the measures of structural variability and to designed entropy, MD RSA and MD iWCN consistently show stronger correlations with sequence entropy. Note that results for MD iWCN are largely identical to those for MD iCN, so only MD iWCN was included here. 47

3.3 Spearman correlations of sequence entropy with MD-derived and crystal-structure derived structural measures. The vertical axes in all plots represent the Spearman correlation of sequence entropy with one structural variable obtained from $15ns$ of molecular dynamics (MD) simulations. The horizontal axes represent the Spearman's rank correlation coefficient of sequence entropy with the same structural variable as in the vertical axes but measured from protein crystal structures. Each dot represents one correlation coefficient for one protein structure. The quantities iCN, iWCN, and RSA have nearly identical predictive power for sequence entropy regardless of whether they are derived from MD simulations or from crystal structures. By contrast, MD RMSF yielded very different correlations than did CS RMSF. 48

3.4 Spearman correlations of sequence entropy with measures of structural variability. Vertical and horizontal axes represent Spearman correlations of the indicated quantities. Each dot represents one correlation coefficient for one protein structure. MD RMSF, CS RMSF, and B factors all explain different amounts of variance in sequence entropy for different proteins. 50

3.5 Spearman correlations of structural quantities with sequence entropy and with the evolutionary rate ratio ω. Nearly all points fall below the $x = y$ line, indicating that structural quantities generally predict as much as or more variation in sequence entropy than in ω. 52

3.6 Principal Component (PC) Regression of sequence entropy against structural variables. **(A)** Variance in entropy explained by each principal component. For most proteins, PC1 and PC3 show the strongest correlations with sequence entropy. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. **(B)** and **(C)** Composition of the three leading components. Red arrows represent the loadings of each of the structural variables on the principal components; black dots represent the amino acid sites in the PC coordinate system. The variables RSA, iWCN, MD RMSF, and Var(χ_1) load strongly on PC1 and weakly on PC2, while B factor and designed entropy load strongly on PC2 and weakly on PC1. 55

3.7 Principal Component (PC) Regression of sequence entropy against the structural variables, including CS RMSF. **(A)** Variance in entropy explained by each principal component. For most proteins, PC1 and either PC2 or PC3 show the strongest correlations with sequence entropy. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. **(B)** and **(C)** Composition of the three leading components. Red arrows represent the loadings of each of the structural variables on the principal components; black dots represent the amino acid sites in the PC coordinate system. 58

3.8 Principal Component (PC) Regression of ω against the structural variables. **(A)** Variance in ω explained by each principal component. For most proteins, PC1 and PC3 show the strongest correlations with ω. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. **(B)** and **(C)** Composition of the three leading components. Red arrows represent the loadings of each of the structural variables on the principal components; black dots represent the amino acid sites in the PC coordinate system. Note that parts B and C are identical to those shown in Figure 3.6. 59

3.9 Principal Component (PC) Regression of ω against the structural variables, including CS RMSF. **(A)** Variance in ω explained by each principal component. For most proteins, PC1 and either PC2 or PC3 show the strongest correlations with ω. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. **(B)** and **(C)** Composition of the three leading components. Red arrows represent the loadings of each of the structural variables on the principal components; black dots represent the amino acid sites in the PC coordinate system. Note that parts B and C are identical to those shown in Figure 3.7. 60

4.1 An Example 2-dimensional Voronoi diagram for bacteriophage T7 lysozyme (Protein Data Bank ID '1LBA'). The red dots represent the backbone C_α atoms projected on the X–Y plane, used as cell seeds in Voronoi tessellation. 72

4.2 A comparison of the prediction power of different Voronoi cell characteristics about site-specific evolutionary rates (ER). Note that all cell characteristic correlate positively with ER, except sphericity which strongly negatively correlates with ER. 74

4.3 The partial correlation strengths of the same Voronoi cell characteristics with sequence evolutionary rates while controlling for the cell area. 75

4.4 A comparison of the correlation strength of 6 different measures of Weighted Contact Number (WCN) with 6 coordinate-independent structural or sequence properties for 209 proteins in dataset. The contact numbers, WCN, are calculated using 6 sets of atomic coordinates: *SC, AA, CB, CA, N, C, O*, used as different representations of individual sites in proteins. The two labels *SC* & *AA* stand respectively for the geometric average coordinates of the Side Chain (SC) atoms and the entire Amino Acid (AA) atoms, excluding hydrogens. 80

4.5 A comparison of the correlation strength of 6 different measures of Voronoi cell areas with 6 coordinate-independent structural or sequence properties for 209 proteins in dataset. The Voronoi cells are generated using 6 sets of atomic coordinates: *SC, AA, CB, CA, N, C, O*, used as different representations of individual sites in proteins. The two labels *SC* & *AA* stand respectively for the geometric average coordinates of the Side Chain (SC) atoms and the entire Amino Acid (AA) atoms, excluding hydrogens. 81

4.6 A comparison of the correlation strength of 6 different measures of B factor with 6 coordinate-independent structural or sequence properties for 209 proteins in dataset. Shown on the horizontal axes, are the 6 representative atomic B factors: *SC, AA, CB, CA, N, C, O* used as flexibility measures of individual sites in proteins. The two variables *SC* & *AA* stand respectively for the average B factor of all Side Chain (SC) atoms and the entire Amino Acid (AA) atoms, excluding hydrogens. 82

4.7 A comparison of the prediction power of five structural variables about site-specific evolutionary rates (ER). All structural quantities correlate positively with ER, with the exception of Weighted Contact Number (WCN) which correlates negatively. For better illustration however, the Spearman's correlation coefficient (ρ) of the inverse of WCN with ER are shown in the Figure. Note that the Spearman's ρ is a rank correlation coefficient, meaning that the use of inverse WCN only changes the sign and not the magnitude of ρ. The abbreviation *SC* refers to the use of average Side-Chain coordinates or average Side-Chain B factor wherever used, and *CA* refers to the use of backbone C_α atomic coordinates for representation of individual sites in proteins. The paired t-test for the significance of the the difference in the observed distributions of correlation strengths are available online in the repository of the project. 85

4.8 A comparison of the prediction power of five structural variables about site-specific Sequence Entropy (SE). All structural quantities correlate positively with SE, with the exception of Weighted Contact Number (WCN) which correlates negatively. For better illustration however, the Spearman's correlation coefficient (ρ) of the inverse of WCN with ER are shown in the Figure. Note that the Spearman's ρ is a rank correlation coefficient, meaning that the use of inverse WCN only changes the sign and not the magnitude of ρ. The abbreviation *SC* refers to the use of average Side-Chain coordinates or average Side-Chain B factor wherever used, and *CA* refers to the use of backbone C_α atomic coordinates for representation of individual sites in proteins. 86

4.9 General behavior of Voronoi cell characteristics versus normalized site-specific evolutionary rates among all sites in all 209 proteins in dataset. The red curves in each plot is obtained by adjacent-averaging of every 3000 sites. The black & orange curves represent respectively the general behaviors of closed & open Voronoi cell characteristics. The blue-shaded area in each plot is a heat map indicating the overall concentration of 75755 sites in all 209 proteins along the horizontal axis. 88

4.10 General behavior of site-specific structural characteristics versus site-specific evolutionary rates among all sites in all 209 proteins in dataset. The red curves in each plot is obtained by adjacent-averaging of every 3000 sites. The blue-shaded area in each plot is a heat map indicating the overall concentration of 75755 sites in all 209 proteins along the horizontal axis. 89

4.11 The scaling behavior of protein maximum extent as defined by Eqn. 4.3 with protein volume for 209 monomeric enzymes in the dataset. The red line is the linear Deming regression fit to logarithms of the two variables with a slope of $D \simeq 2.47 \pm 0.06$. 92

4.12 The scaling behavior of protein's radius as defined by Eqn. 4.4 with protein length for 209 monomeric enzymes in the dataset. The mean & median length of the proteins are 362 & 315 respectively. The red line is the linear Deming regression fit to logarithms of the two variables with a slope of $D \simeq 2.60 \pm 0.08$. 93

4.13 An illustration of the strong positive correlation of X-ray crystallography resolution with the ratio of the backbone C atomic B factor to the average amino acid B factor ($\mathrm{BF}_C/\mathrm{BF}_{AA}$), averaged over all sites in individual proteins, highlighting the significant contributions of noise and model errors to atomic B factor values. The Spearman's correlation coefficient between the two quantities is $\rho \sim 0.76$. No significant correlation would be expected in the absence of noise due to limited resolution of the X-ray crystallography of proteins. Each filled circle in the plot represents one protein in the dataset of 209 enzymes used in this work. 97

4.14 The average absolute Spearman's correlation strengths of the Weighted Contact Number with power-law kernel as given by Eqn. 4.1 for different values of the free parameter of the kernel α. The solid black line represents the mean correlation strength in the entire dataset of 209 proteins, and the dashed black line indicates the median of the distribution. The green-shaded region together with the two read dashed lines represent the 25% & 75% quartiles of the correlation strength distribution. Note that for $\alpha > 0$ the sign of the correlation strength ρ is the opposite of the sign of ρ for $\alpha < 0$. In addition ρ is undefined at $\alpha = 0$ and not shown in this plot. The parameter values at which the Spearman's correlation coefficient reaches the maximum over the entire dataset are given in Table 4.1. 102

4.15 The average absolute Spearman's correlation strengths of the Weighted Contact Number with Gaussian kernel as defined by Eqn. 4.16 for different values of the free parameter of the kernel σ. The solid black line represents the mean correlation strength in the entire dataset of 209 proteins, and the dashed black line indicates the median of the distribution. The green-shaded region together with the two read dashed lines represent the 25% & 75% quartiles of the correlation strength distribution. The parameter values at which the Spearman's correlation coefficient reaches the maximum over the entire dataset are given in Table 4.1. 104

4.16 The average absolute Spearman's correlation strengths of the Weighted Contact Number with exponential kernel as defined by Eqn 4.16 for different values of the free parameter of the kernel (the exponential mean λ). The solid black line represents the mean correlation strength in the entire dataset of 209 proteins, and the dashed black line indicates the median of the distribution. The green-shaded region together with the two read dashed lines represent the 25% & 75% quartiles of the correlation strength distribution. The parameter values at which the Spearman's correlation coefficient reaches the maximum over the entire dataset are given in Table 4.1. 105

4.17 The average absolute Spearman's correlation strengths of the Weighted Contact Number with hard-sphere cutoff kernel for different values of the free parameter of the kernel (R_C). This definition of WCN measures of the number of amino acids within a spherical neighborhood of radius R_C around a given amino acid in the site of interest. The solid black line represents the mean correlation strength in the entire dataset of 209 proteins, and the dashed black line indicates the median of the distribution. The green-shaded region together with the two read dashed lines represent the 25% & 75% quartiles of the correlation strength distribution. The parameter values at which the Spearman's correlation coefficient reaches the maximum over the entire dataset are given in Table 4.1. 106

5.1 A comparison of the strength of the Spearman's correlation strength of sequence evolutionary rates (r4sJC) with side chain Weighted Contact Number (on the vertical axes of plots) vs. correlation strengths of other structural properties with evolutionary rates (on the horizontal axes). Detailed description of the structural properties is given are given Chapters 2 & 4. The red lines in each plot represent equality. It is evident from all plots that for any given protein in dataset, the correlation strength of one structural property is a good proxy measure of the correlation strength of any other structural property with sequence variability measures. The correlation strengths of the two correlation measures on the vertical and horizontal axes are provided on the bottom-right of each plot. 111

5.2 A comparison of the strength of the Spearman's correlation strength of sequence entropy with side chain Weighted Contact Number (on the vertical axes of plots) vs. correlation strengths of other structural properties with sequence entropy (on the horizontal axes). Detailed description of the structural properties is given are given Chapters 2 & 4. The red lines in each plot represent equality. It is evident from all plots that for any given protein in dataset, the correlation strength of one structural property is a good proxy measure of the correlation strength of any other structural property with sequence variability measures. The correlation strengths of the two correlation measures on the vertical and horizontal axes are provided on the bottom-right of each plot. 112

5.3 Hierarchical clustering diagram of the Spearman's correlation matrix of all pdb-level properties considered in this work, used to identify groups of closely related variables that potentially represent a similar underlying property of proteins. A full size of the diagram and the meaning of each of the variables are available in the permanent online repository of the work (c.f., Section 5.2). 119

5.4 **The Spearman correlation matrix for the strongest sequence–structure correlation (denoted by** $r.r4s.wcn$**) and the prominent determinants of the strengths of this relation**. The variables on the diagonal elements of the matrix from top to bottom represent respectively, the strongest sequence–structure relation – i.e., the absolute Spearman's correlation of evolutionary rates (r4s) with side-chain Weighted Contact Number (WCN) – followed by important protein properties that appear to modulate the strength of this relation: variance of sequence entropy (*sd.se*), variance of site-specific evolutionary rates (*sd.r4s*), variance of back-bone hydrogen bond energies (*sd.hbe*), and the fraction of amino acids in helical & $\beta-$sheet secondary structures in the protein (*mn.helix* & *mn.betas* respectively). 120

5.5 **Sequence–structure correlation strength versus sequence divergence.** The plot illustrates the relationship between the strength of a representative sequence–structure correlation (sequence entropy – Weighted Contact Number) and the sequence divergence as measured by the variance of protein sequence entropy. The black circles represent 209 proteins used in this work. For comparison, the red circles represent data from 9 viral proteins taken from Chapter 3 [104]. 124

Chapter 1

Introduction

Proteins are known as long chains of amino acids tightly packed in almost unique three dimensional conformations. The *native structure* of a protein, generally determined by Nuclear Magnetic Resonance (NMR) or X-ray crystallography methods, is believed to correspond to the global minimum free energy conformation. Studies on protein dynamics however, have revealed a highly rugged hierarchical energy landscape for proteins, composed of sets of many small energy barriers, each residing in a local minimum energy basin [20, 26]. The rugged landscape is primarily a result of the large number of degrees of freedoms that the backbone and side-chain atoms of amino acids possess in polypeptides. In this picture of the energy landscape, different minima correspond to changes in the relative orientations of the secondary structures (i.e., α-helices and β-sheets) coupled with primary structure (i.e., side-chain) rearrangements on timescales $\sim 1 - 10\ ns$, such that the close packing of the protein interior is conserved. Consequently, proteins are expected to exhibit flexibility in vivo, as evidenced and observed in Molecular Dynamic (MD) simulations.

Parallel to spatial variations in structure, proteins also exhibit vari-

ability in their amino acid sequences on evolutionary timescales. Significant variations can be observed in the AA sequence among the divergent members of a protein family, while conserving the relative similarities of their native conformations. These evolutionary variations can be due to a combination of point mutations, insertions, deletions or sometimes the rearrangement of the domains in the protein sequence [8].

During the past years, the role of protein structure and dynamics at the residue level on its sequence variation and evolution has gained considerable attention. Several recent works have shown that protein structure can predict site-specific evolutionary sequence variation. In particular, sites that are buried and/or have many contacts with other sites in a structure have been shown to evolve more slowly, on average, than surface sites with few contacts. In Chapter 3 first I present a comprehensive study of the extent to which numerous structural properties can predict sequence variation using a set of viral proteins. The quantities considered include buriedness (as measured by relative solvent accessibility), packing density (as measured by contact number), structural flexibility (as measured by Debye-Waller factors, root-mean-square fluctuations, and variation in protein backbone and side-chain dihedral angles), and variability in designed structures. The structural flexibility measures are obtained both from Molecular Dynamics simulations performed on 9 non-homologous viral protein structures and from variation in homologous variants of proteins in the dataset, where available. Measures of variability in designed structures are obtained from flexible-backbone de-

sign using the Rosetta software. I find that most of the structural properties correlate with site-specific measures of sequence variability in the majority of structures, though the correlations are generally weak, with Spearman's correlation coefficients of $\rho \in [0.1, 0.4]$. Moreover, measures of amino acid buriedness and packing density are found to be better predictors of evolutionary variation than was structural flexibility. Finally, variability in designed structures turns out to be a weaker predictor of evolutionary variability than buriedness or packing density, but it is comparable in its predictive power to the best structural flexibility measures. I conclude that simple measures of buriedness and packing density are better predictors of evolutionary variation than are more complicated predictors obtained from dynamic simulations, ensembles of homologous structures, or computational protein design.

Motivated by the findings of Chapter 3, I expand the study of structure-sequence relationships in chapter 4 to a significantly larger dataset of 209 monomeric enzyme proteins in contrast to the viral dataset considered in the previous chapter. Confirming the findings of Chapter 3, I further present a wider and deeper analysis of site-specific structural characteristics, including measures of local flexibility and packing in proteins. I identify and highlight the potential caveats and biases associated with each of the site-specific structural characteristics, in particular flexibility and density measures. Then I introduce possible remedies to improve or reduce bias in the definitions and estimates of the site-specific structural characteristics of proteins. In particular, I used Voronoi tessellation methods from the fields of Computational Geometry and

Condensed Matter Theory to obtain parameter-free measures of local packing density and less-biased measures of local flexibility in proteins. Contrary to the common representation of protein structure using the coordinates of C_α backbone atoms, I show that the 3-dimensional structures of proteins are best represented by the geometric center of amino acid side-chain coordinates. The use of site-chain coordinates in Elastic Network Models (ENM) of proteins in particular results in significantly better predictions of local residue fluctuations and sequence evolutionary rates, further details of which will be discussed in Chapter 4. Contrary to recent reports, I show that there is no unique best kernel for modelling residue-residue interactions in protein structure, based upon which the Kirchhoff (connectivity) matrix of ENM is constructed. This finding highlights the existence of diverse energy landscapes for proteins and the fact that no single potential-of-mean-force can uniquely describe all interactions between individual sites in proteins.

Finally, in chapter 5 I attempt the results of the search potential modulators of sequence-structure correlation strengths in the dataset of 204 monomeric enzymes studied Chapter 4, which appear to vary widely among different proteins with absolute correlation strengths ranging from 0.1 to 0.8. I discuss the main protein characteristics responsible for the general patterns of protein evolution, and identify sequence divergence as the primary determinant of the strengths of virtually all structure-evolution relationships, explaining $\sim 10 - 30\%$ of the observed variations in sequence-structure correlation strengths. In addition to sequence divergence, several structural character-

istics of proteins are identified that are moderately but significantly coupled with the strength of sequence-structure relations. Specifically, proteins with more homogeneous back-bone hydrogen bond energies, larger fractions of helical secondary structures and less fraction of beta sheets tend to have the strongest sequence-structure relations.

Chapter 2

Site-Specific Structural and Evolutionary Characteristics of Proteins

In this chapter, I introduce and briefly discuss some of the most important site-specific structural and evolutionary characteristics of proteins. These site-specific properties and their interrelationships will be then extensively studied in sets of viral and enzyme proteins in the following chapters. The structural characteristics of proteins can be broadly divided into three main categories: measures of amino acid residue buriedness (or conversely, exposure to solvent molecules), site-specific flexibility, and local packing density. There are also other site-specific characteristics that do not fall necessarily in one of the three aforementioned categories, such as hydrophobicity and the average Hydrogen bond (H-bond) energy of the amino acid occupying a specific site in protein. The hydrophobicity scale in particular differs from the aforementioned site-specific structural characteristics in that it cannot be derived solely from the protein's 3-dimensional structure, without the knowledge of the type of the amino acid occupying individual sites in proteins. Nevertheless, for the sake of comprehensiveness and their potential effects on sequence-structure relations it will be briefly discussed and included in this study. In addition, I describe measures of sequence variability in the following sections that are

obtained from protein design or from the study of the energetics of amino acid variations at individual sites.

2.1 Site-Specific Measures of Sequence Variability

The amino acid variability of a given site in protein can be a strong indicator of its importance in function or the stability of the structure of the protein. Calculation of sequence variability at the amino acid level firstly requires collection of sequence data from all divergent members of the protein family, and subsequently the alignment of the collected sequences. Once sequences are collected and aligned, a very simple measure of variability at the amino acid level can be obtained by calculating the Shannon entropy (H_i), the so-called *sequence entropy* [105], of a given column i in the aligned sequences, based on the assumption that the occurrence of each of the 20 amino acids is equally likely at any given site in the alignments:

$$H_i = -\sum_j P_{ij} \ln P_{ij} \tag{2.1}$$

in which P_{ij} is the relative frequency of amino acid j at position i in the alignment. It should be noted that sequence entropy is not an exact equivalent of site-specific evolutionary rates, although the two quantities often correlate strongly with each other (e.g., Figure 4.10 in Chapter 4).

A more accurate measure of sequence variability requires construction of phylogenetic trees that provide a chronological connection among the

aligned sequences in order to calculate the evolutionary rates. The evolutionary rate, by definition, is the average number of mutations accumulated by by individual sites in diverging sequences over evolutionary timescale. For a constant evolutionary rate, the distance between two sequences, defined as the expected number of substitutions per site, will increase linearly with the time of divergence. A simple measure of distance in this scenario would be the proportion of different sites between the two sequences, often called the *p distance*. This definition is however too simplistic for highly dissimilar sequences with $p \gtrsim 5\%$. For example, a variable site may be the result of more than one substitution and an apparently non-variable site could be a result of back or parallel substitutions. This implies that for highly diverged sequences p is not a linear function of evolutionary time. A correct estimate of the substitutions thus requires a probabilistic model to take into account multiple substitutions for each site.

Multiple methods of defining evolutionary units exist: The unit of evolution can be defined at the DNA level using the nucleotides, or using amino acids in protein sequence, or using the nucleotide triplets of coding sequences. The latter is particularly more appropriate for the study of protein evolution, since DNA-based methods are incapable of capturing the codon redundancy in the coding sequences of amino acids, while methods based on amino acids as units of evolution omit important fine details, such as synonymous vs. non-synonymous substitutions that can only be revealed through the analysis of coding sequences. The redundancy in the number of codons that code for

each amino acid is best captured by codon-based evolutionary models. This redundancy results in the majority of mutations in the coding sequence to be *silent* or *synonymous*, causing no change in the structure of protein, while less frequent *nonsynonymous* mutations can cause a change in the type of the amino acid, and potentially in the three-dimensional structure of protein.

Such a distinction between the two types of codon substitutions, allows one to define an intuitive measure of selection pressure on different amino acid sites in a protein [54]. In this case, the *evolutionary rates ratio* (ω) as a measure of selective pressure on a given site in the aligned sequence data is often quantified via the ratio of the number of non-synonymous substitutions per non-synonymous site ($\mathrm{d}N$) to the number of synonymous substitutions per synonymous site ($\mathrm{d}S$),

$$\omega \equiv \frac{\mathrm{d}\,N}{\mathrm{d}\,S}, \tag{2.2}$$

which is an indicator of selective pressure acting on a protein-coding gene [44]. A synonymous substitution refers to the evolutionary substitution of one nucleotide base with another in the codon sequence of the protein, such that the resulting amino acid in the specific site of interest in protein is not altered, whereas a non-synonymous nucleotide substitution in the codon sequence of protein alters the amino acid sequence of the protein in the specific site of interest. A completely neutral evolution would correspond to $\omega \sim 1$, while $\omega < 1$ indicates purifying selection and $\omega > 1$ implies positive selection.

2.1.1 Codon models of evolutionary rates

A wide variety of substitution models and software already exist for the purpose of sequence alignment, construction of evolutionary trees, and finally the estimation of evolutionary rates (ω), some of which will be discussed and used in the following chapters. As discussed in the previous paragraphs, the unit of evolution in a codon model is the codon triplet rather than single nucleotide or amino acid as in the DNA or amino acid evolutionary models.

The estimation of the number of substitutions in the coding sequences of proteins requires building a probabilistic model to describe the changes between codons. In this regard, substitutions at any given site are commonly described by continuous-time Markov chain models. The use of Markov chain comes with a presumption that the substitution from state i to another state j depends only on the two states i & j and not the past history of states, a characteristic of Markovian processes. In such a model, the changes of the coding sequence in time depends only on the current state of the sequence and not on how the current state has been reached from the past. Denoting the instantaneous rate of change from codon i to j among all 61 possibilities by q_{ij}, the codon substitution is modelled by the 61×61 substitution-rate matrix $Q = \{Q_{ij}\}$, so that the quantity $q_{ij}\Delta t$ gives the probability that any given codon i will change to a different codon j in the small time interval Δt. The *transition-probability matrix*, $P(t) = \{p_{ij}\}$, can be then obtained from the substitution-rate matrix using the differential equation,

$$\frac{\mathrm{d}\,P(t)}{\mathrm{d}\,t} = P(t)Q, \tag{2.3}$$

with the boundary condition $P(t = 0) = I$ with I representing the identity matrix [35]. This differential equation has the solution $P(t) = \exp^{Qt}$ [13].

Markov-chain models of codon substitution were first proposed by [32, 80] and later expanded by [31], in which the substitution of codon triplet by another is described as a Markovian process. In its simplest form, the elements of the substitution-rate matrix Q_{ij} according to the model of Goldman & Yang (1994) [32] can be written as [129],

$$q_{ij} = \begin{cases} 0, & \text{if } i \;\&\; j \text{ differ at two or three codon positions.} \\ \pi_j, & \text{if } i \;\&\; j \text{ differ by a synonymous transversion.} \\ \kappa\pi_j, & \text{if } i \;\&\; j \text{ differ by a synonymous transition.} \\ \omega\pi_j, & \text{if } i \;\&\; j \text{ differ by a nonsynonymous transversion.} \\ \omega\kappa\pi_j, & \text{if } i \;\&\; j \text{ differ by a nonsynonymous transition.} \end{cases} \tag{2.4}$$

in which κ is the transition/transversion ratio, ω is the ratio of nonsynonymous to synonymous substitutions, and π_j is the equilibrium frequency of the j^{th} codon. While the parameter ω describes the selection at the amino acids level in protein, the two other quantities, π_j & κ describe the mutational process at the nucleotide level in DNA.

Alternatively, Muse and Gaut (1994) [80], suggested a substitution-rate model defined as,

$$q_{ij} = \begin{cases} 0, & \text{if } i \ \& \ j \text{ differ at two or three codon positions.} \\ \alpha\pi_{jn}, & \text{if } i \ \& \ j \text{ differ by a synonymous substitution.} \\ \beta\pi_{jn}, & \text{if } i \ \& \ j \text{ differ by a nonsynonymous substitution.} \end{cases} \tag{2.5}$$

in which α & β represent the synonymous & nonsynonymous substitution rates respectively, and π_{jn} stands for the equilibrium frequency of the n^{th} nucleotide of codon j.

Comparing the two models described above, one can notice that the highly simplified model of Muse and Gaut (1994) does not correct for the transition/transversion bias, unlike the model of Goldman & Yang (1994). The latter model is therefore more accurate. The improved accuracy however comes at the cost of more computation and the requirement of having large sequence sample to avoid degeneracies in parameter estimations of the model.

The estimation of the free parameters of the model (e.g., the variables ω in Equation 2.4) is typically done by first constructing the likelihood function of the model. Given the sequence alignment data and the phylogenetic tree of the species, the best parameters of the model are then obtained by maximizing the likelihood function. Throughout the following chapters, the model of Goldman & Yang (1994) will be adopted to estimate the evolutionary rates ratio ω.

2.1.2 Amino acid models of evolutionary rates

An alternative approach to evolutionary rate estimation besides codon substitution models is the amino acid based method, a prime example of which

is implemented in *rate4site* software by Pupko et al. (2002) [88]. Rate4site estimates the rates of evolution of amino acid sites using the maximum likelihood method by considering the topology and branch lengths of the phylogenetic tree in addition to the underlying stochastic processes. The branch lengths of the phylogenetic tree represent the average evolutionary rate across all sites, and the site-specific evolutionary rate, r, indicates how fast this site evolves relative to the average rate across the entire sequence. Thus a rate of 2.0 would indicate a site that is evolving two times faster than the rest of the sequence on average.

The rat4site algorithm obtains the rate parameter r_j for the j^{th} site using a maximum likelihood approach similar to that of [127] for modelling sequence evolution, with the only difference that here rates are estimated for individual sites. The higher the variability of the site j, the higher the value of the rate r_j will be. To expand on this, consider an example four-taxon unrooted phylogenetic tree as in Figure 2.1. Given the tree T characterized by the tree topology τ and the associated branch lengths t, rate4site calculates the likelihood of observing data given r & T as,

$$\begin{aligned} P\big(\text{data}|r,T\big) &= \sum_{X_1,X_2\in\{20\text{ Amino-Acids}\}} \pi_{X_1} \times P_{X_1,M}\big(r.t_1\big) \\ &\times \; P_{X_2,G}\big(r.t_2\big) \times P_{X_2,M}\big(r.t_3\big) \\ &\times \; P_{X_1,I}\big(r.t_4\big) \times P_{X_1,X_2}\big(r.t_5\big) \end{aligned} \tag{2.6}$$

in which π_{X_1} represents the frequency of the amino acid X_1, $P\big(X1,X2\big)$ is

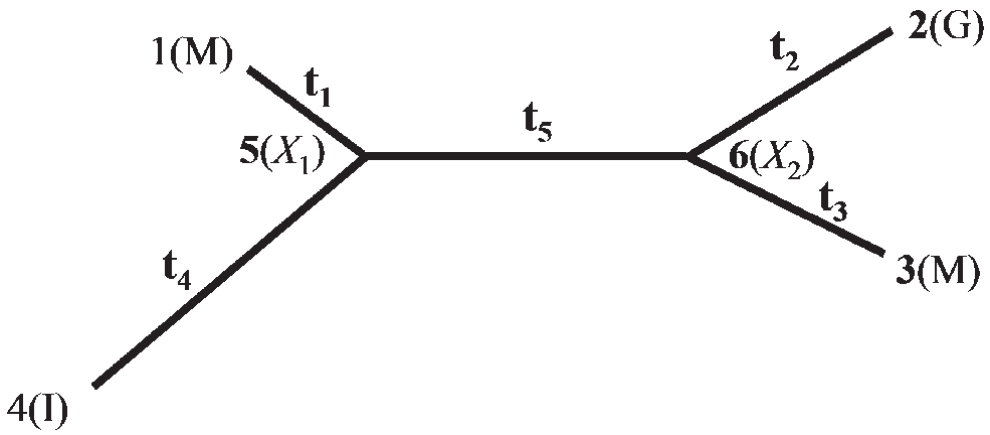

Figure 2.1: An example four-taxon unrooted tree taken from [72] for illustration purposes. The external nodes (i.e., leaves) are labelled from 1 to 4 while internal nodes are 5 & 6. Branch lengths are denoted by t_i and the capital letters in parentheses on each node represent the one-letter abbreviations for the amino acids.

the probability of the replacement of the amino acid X_1 by the amino acid X_2 along a branch of length t, given the evolutionary rate of r at the site of interest. The best value of r for each site is then obtained by maximizing the likelihood function $P\big(\text{data}|r,T\big)$ according to 'postorder tree traversal' algorithm of Felsenstein (1981) [21].

2.2 Site-Specific Sequence Variability as Measured from Protein Design

In addition to site-specific evolutionary rates and sequence entropy, an independent proxy measure of sequence variability can be also obtained purely from computational protein design without recourse to natural sequence data collection and alignment. The world of computational protein design has witnessed tremendous growth and popularity over the past decade, primarily due to its promise for novel drug design and engineering protein folds that have not been observed in nature. Protein design software, such as RoswttaDesign, have been already developed over the past two decades that are capable of generating protein sequences that fold stably into pre-specified structures. The designed sequences can be then aligned and compared to alignments of natural protein sequences to assess how closely the designed sequences correlate with patterns of variability found in natural protein sequences.

2.3 Site-Specific Stability Contribution to Protein Native Conformation

A measure of thermodynamic stability changes due to amino acid substitutions at individual sites in proteins can be defined and obtained following the stability threshold model of Bloom et al. (2005) [6], which was also recently further studied by Echave et al. (2014) [19]. Suppose the required change in the Gibbs free energy of a protein upon folding to the native conformation is ΔG_{native}. According to Bloom's model, there is a certain energy threshold $\Delta G_{thresh} > \Delta G_{native}$, such that any protein conformation corresponding to a Gibbs free energy, $\Delta G > \Delta G_{thresh} > \Delta G_{native}$ is significantly different from the native conformation of the protein and therefore unstable or biophysically unimportant. Now, any individual amino acid substitution in the native conformation of the protein may result in a change in the native Gibbs free energy change of the protein of the amount,

$$\Delta\Delta G = \Delta G - \Delta G_{native}. \tag{2.7}$$

Defining $\Delta\Delta G_{thresh} = \Delta G_{thresh} - \Delta G_{native} > 0$, now any point mutation in protein sequence with a change in the free energy of the protein $\Delta\Delta G > \Delta\Delta G_{\text{thresh}}$, would result in a new conformation of the protein which would be functionally disruptive. Therefore, such amino acid substitution is expected to occur rarely in protein sequence on evolutionary timescales. Stability changes due to amino acid substitutions are expected to differ from one site in protein to another, and hence different sites in proteins are expected to have different

tolerance to amino acid substitutions. Under the assumption of symmetric mutations (i.e., single site mutations obeying the principle of detailed balance), Echave et al. (2014) define a measure of the stability of the protein upon random point mutations, details of which is given in [19]. Hereafter throughout the rest of the work, this quantity is denoted by $\Delta\Delta G$ *rate*.

The $\Delta\Delta G$ rate estimates for all structures in the following chapters are calculated using data from FoldX software [103]. A low value of $\Delta\Delta G$ rate for a given site in protein indicates a high chance of structure perturbation upon substitution and therefore the amino acid in the site of interest is expected to be highly conserved on evolutionary timescales.

2.4 Site-Specific Solvent-Accessible Surface Area

The Solvent Accessible Surface Area (SASA) of amino acids in proteins has emerged as a very popular tool in Biochemistry and Computational Biology over the past few decades. This quantity measures the surface area of an amino acid in a given site in protein that is exposed to solvent molecules (typically water) surrounding the protein (Figure 2.2). It is therefore expected that amino acids buried in protein core would in general have less SASA values in contrast to sites near the surface of the protein. The SASA for each amino acid in protein is basically measured by rolling a ball of an effective radius of the solvent molecule on the surface of the protein. A variety of software have been developed over the past four decades for the calculation of site-specific SASA in proteins. A popular software in this regard is DSSP [48] which can calcu-

late SASA for individual amino acids in all sites in proteins using a spherical probe of radius $\sim 1.5\mathring{A}$, representing a water molecule. Since the 20 naturally occurring amino acid molecules come in different sizes, it is necessary to normalize the SASA values of individual amino acids to the their corresponding *maximum solvent accessibility*. The maximum SASA has been traditionally obtained from experimental measurements [95]. Alternatively, the SASA values from DSSP can be instead normalized to the computationally calculated maximum SASA values of [118] to obtain the Relative Solvent Accessibility (RSA) for all individual sites in all proteins. Figure 2.3 illustrates the general behavior of protein sequence evolution and its relation to solvent accessibility of amino acids.

2.5 Site-Specific Flexibility and Fluctuation Measures in Proteins

The local flexibility of proteins at the amino acid level can be estimated by several independent methods, such as thermal atomic fluctuation measurements from the X-ray crystallography of proteins, or the Root-Mean-Squared Fluctuations (RMSF) as measured from Molecular Dynamics (MD) simulations, or simply from structural superposition of homologous proteins. The former is among the most popular proxy measures of site-specific flexibility due to its simplicity and accessibility from Protein Data Bank (PDB) files, whereas RMSF from MD simulations are normally computationally very expensive to calculate. In the following two subsections, each of the two local

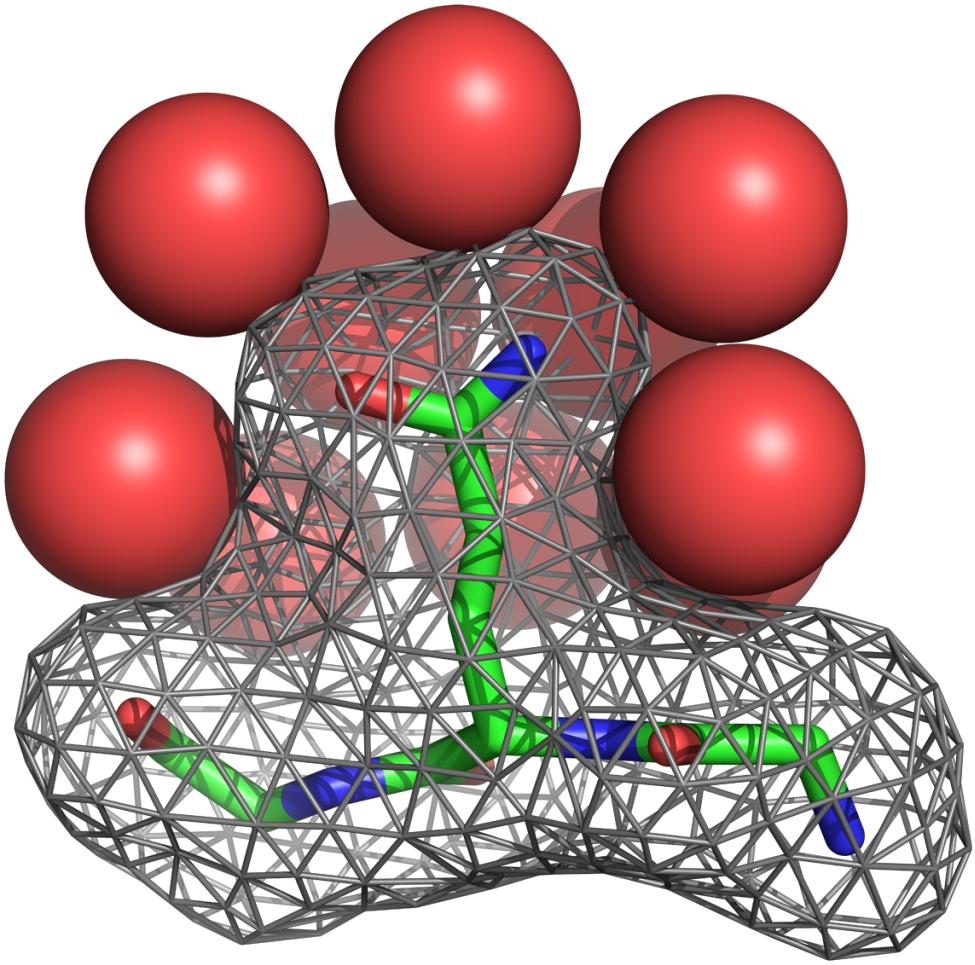

Figure 2.2: An illustration of methodology that is typically used for the calculation of Solvent Accessible Surface Areas of amino acids in individual sites in proteins. Depicted in this figure is the Glutamine molecule surrounded by solvent molecules (typically water) represented by the red spheres. For better illustration, the solvent molecules in front of the Glutamine have been removed. An approximate measure of solvent accessibility can be obtained by counting the number of spherically-shaped solvent molecules of radius $\sim 1.5\mathring{A}$ that can fit around an amino acids in a given site in protein. The solvent accessibility is therefore a discrete quantity by definition. (Illustration is courtesy of Austin G. Meyer, e.g., [118])

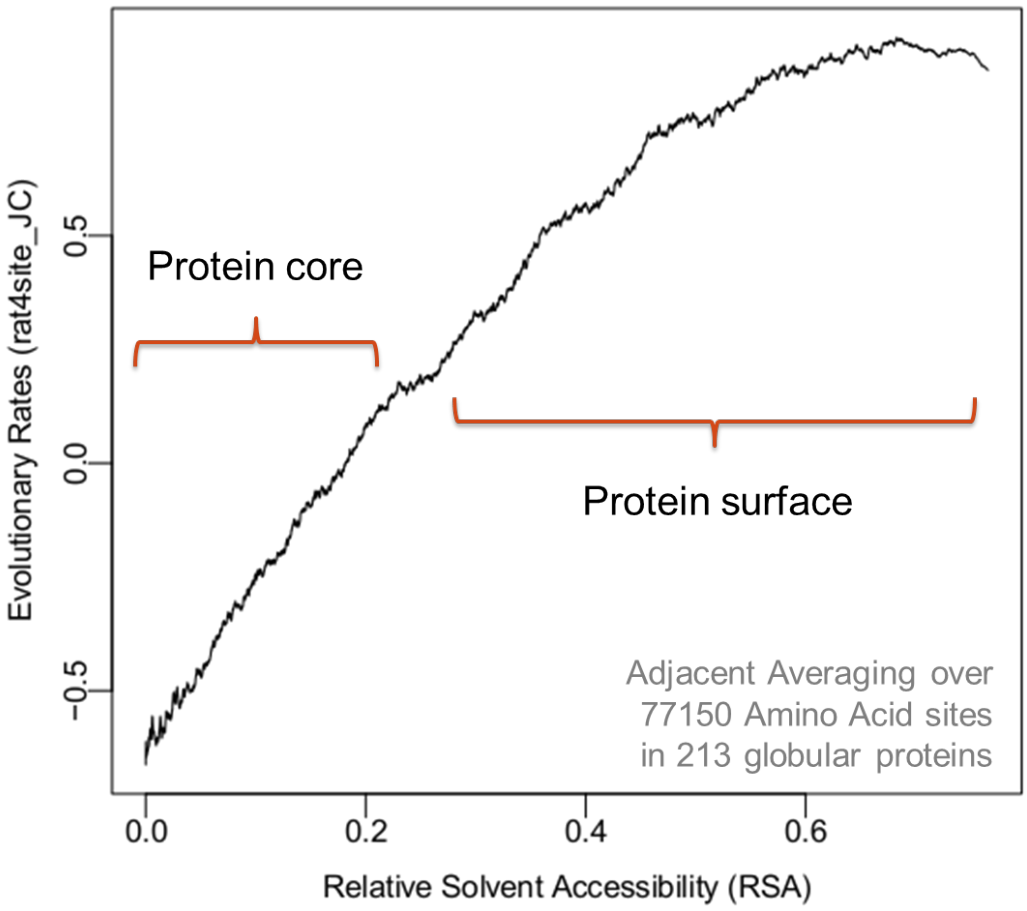

Figure 2.3: A general positive trend between the Relative Solvent Accessibility of an amino acid in protein and its sequence evolutionary rates is normally seen in all proteins. Amino acids with RSA $\lesssim 0.2$ are considered to be buried deep in the core of protein, whereas sites with RSA$\gtrsim 0.2$ are considered to be part of the surface of the protein. The average curve shown in the plot was obtained by adjacent averaging over all sites in a dataset of 213 monomeric enzymes (c.f., Chapter 4).

flexibility measures will be briefly introduced and discussed.

2.5.1 Thermal Atomic Fluctuations as Proxy Measures of Amino Acid Flexibility in Proteins

A popular proxy measure of local flexibility or fluctuation in different parts of a crystalline structure, in particular proteins on the atomic scale, is a quantity called *B factor* or the *temperature factor*. Atoms with low values of B factor belong to parts of the protein structure that is well-ordered, whereas atoms with high values of B factor belong to parts of the protein structure that is more flexible. The B factors for all atoms in proteins are generally obtained from X-ray crystallography and are deposited in the Protein Data Bank files. For each atom in protein, it can be calculated from the DebyeWaller factor (DWF), named after Peter Debye and Ivar Waller [15,120], which is often used in Condensed Matter Physics to describe the attenuation of x-ray scattering caused by thermal motion of the scattering object. For a given scattering vector $\mathbf{q}$ and a scattering center displacement vector $\mathbf{u}$, DWF is expressed as,

$$\mathrm{DWF} = \langle \exp(i\mathbf{q}\cdot\mathbf{u})\rangle^2 = \exp\left(-\frac{q^2\langle u^2\rangle}{3}\right), \tag{2.8}$$

in which the second equality holds only under the assumption of a harmonic and isotropic potential in which the scattering center fluctuates. Here the variable u represents the magnitude of the displacement vector $\mathbf{u}$ and,

$$q = \frac{4\pi \sin\left(\frac{\theta}{2}\right)}{\lambda}, \tag{2.9}$$

represents the magnitude of the scattering vector at angle θ with respect to an incident wave of wavelength λ. In X-ray crystallography of proteins, the quantity B factor is defined as,

$$\mathrm{BF} = 8\pi^2\langle u^2\rangle. \tag{2.10}$$

Although, B factor is an atomic measure of flexibility and fluctuation in proteins, it has become a very popular proxy measure of amino acid flexibility in the studies of protein dynamics and benchmarking of different Elastic Network Models of proteins. In this regard, the flexibility of a residue is often represented by the B factor of C_α backbone atom of the amino acids in proteins.

2.5.2 Site-Specific Fluctuations from Protein Conformational Ensemble

Alternative measures of amino acid and side-chain fluctuations can be also obtained computationally via Molecular Dynamics simulation of proteins in an aqueous environment similar to the physiological environment of proteins in vivo, or from the alignment of homologous crystalline structures. A prime example of such quantity is RMSF. For each Cα atom in protein, RMSF is calculated based on an ensemble of protein conformations extracted from MD trajectories or homologous crystalline structures. For MD trajectories, the conformational snapshots of the protein structure are first fit to a reference structure, generally the protein crystalline structure or the average structure

over all MD snapshots. This fitting removes any translational or rotational motion of the entire protein structure so that local fluctuation could be captured with more accuracy. The quantity RMSF can be then calculated as,

$$\mathrm{RMSF}_j = \Big[\sum_i \big(\mathbf{r}_i^{(j)} - \mathbf{r}_0^{(j)} \big)^2 \Big]^{1/2} \tag{2.11}$$

where RMSF_j is the root-mean-square fluctuation at site j, $\mathbf{r}_i^{(j)}$ is the position of the Cα atom of residue j at MD frame i, and $\mathbf{r}_0^{(j)}$ is the position of the Cα atom of residue j in the original crystal structure.

To calculate RMSF from homologous structures, the structures are similarly first aligned, based upon which RMSF is calculated as,

$$\mathrm{RMSF}_j = \Big[\sum_i w_i \big(\mathbf{r}_i^{(j)} - \langle \mathbf{r}^{(j)} \rangle \big)^2 \Big]^{1/2}, \tag{2.12}$$

where $\mathbf{r}_i^{(j)}$ now stands for the position of the Cα atom of residue j in structure i, $\langle \mathbf{r}^{(j)} \rangle$ is the mean position of that Cα atom over all aligned structures, and w_i is a weight to correct for potential phylogenetic relationship among the aligned structures.

Compared to B factor, the computational expense of MD simulations and the availability of homologous crystal structures can severely restrict the applicability of RMSF as a measure of site-specific fluctuation and flexibility. Furthermore, the presence of collective secondary structure motions can potentially introduce significant biases in RMSF estimates. This is primarily due

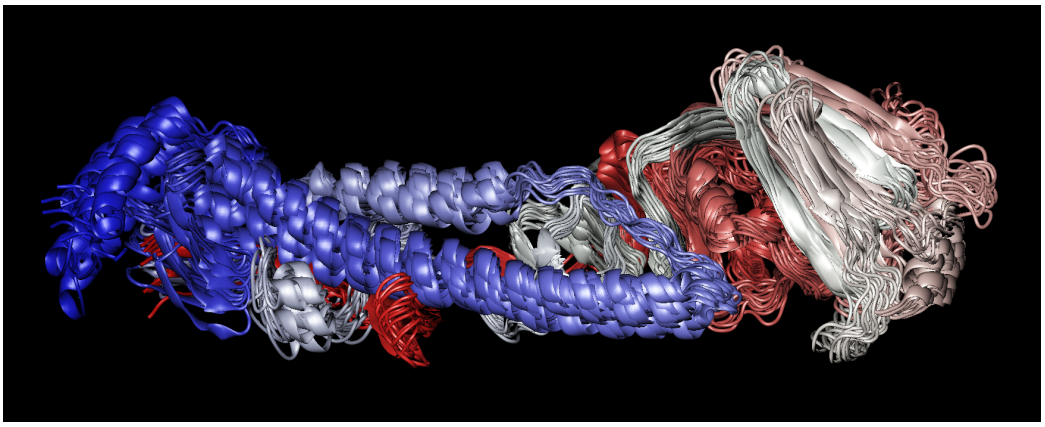

Figure 2.4: A cartoon representation of Influenza Hemagglutinin protein *1RD8_AB*, illustrating the collective motions in secondary structures (i.e., α-helices and β-sheets). Color coding begins from the start (N-terminus) of the sequence in blue to the end (C-terminus) in red. The figure is a result of superposition of 12 conformational snapshots obtained every $100ps$ from Molecular Dynamics simulation of 1RD8_AB. The collective motion of amino acids in secondary structures, can potentially introduce strong biases in estimations of RMSF of amino acids in individual sites.

to the dependence of RMSF definition on the set of external coordinates of individual sites in proteins. An example dominant effect of collective motions in protein structure on the local fluctuation measurements is illustrated in Figure 2.4 where multiple snapshots of chains A & B of the Influenza Hemagglutinin protein trimer (*1RD8_AB*) are superimposed on each other.

By contrast, the variabilities in the dihedral angles of amino acid side-chains, in particular χ_1, appear to be ideal proxy measures of local amino acid flexibility, since unlike root-mean-square fluctuation measures, their calculations does not involve the use of an external reference point in Cartesian coordinates of the protein and therefore, the amount of systematic bias in-

troduced in fluctuation estimates is minimized. In addition, the variability in dihedral angles of amino acids represent the local fluctuation of the amino acid as a whole molecule in a given site, and are therefore better representations of site-specific flexibility compared to atomic B factors. This accuracy however, comes at the cost of extensive computational resources required to perform long-trajectory MD simulations on timescales of $1 - 100ns$. For small protein dataset, the computational cost maybe within the realm of current computational technologies, an example of which will be presented in Chapter 3.

To assess variability in backbone and side-chain dihedral angles, the quantities $\mathrm{Var}(\phi)$, $\mathrm{Var}(\psi)$, and $\mathrm{Var}(\chi_1)$ can be used, where ϕ and ψ refer to backbone dihedral angles and χ_1 is the first dihedral angle in the amino acid side-chain. The variance of a dihedral angle can be defined according to the most common definition in directional statistics: First, a unit vector $\mathbf{x}_i$ is assigned to each dihedral angle α_i in the sample. The unit vector is defined as $\mathbf{x}_i = (\cos(\alpha_i), \sin(\alpha_i))$. The variance of the dihedral angle is then defined as,

$$\mathrm{Var}(\alpha) = 1 - ||\langle \mathbf{x} \rangle||, \tag{2.13}$$

where $||\langle \mathbf{x} \rangle||$ represents the length of the mean $\langle \mathbf{x} \rangle$, calculated as $\langle \mathbf{x} \rangle = \sum_i \mathbf{x}_i / n$. Here, n is the sample size. The variance of a dihedral angle is, by definition, a real number in the range $[0, 1]$, with $\mathrm{Var}(\alpha) = 0$ corresponding to the minimum variability of the dihedral angle and $\mathrm{Var}(\alpha) = 1$ to the maximum, respectively [4]. Since the χ_1 angle is undefined for Ala and Gly, all sites

in protein containing these two amino acids must be consequently discarded in the analyses and studies involving χ_1.

2.6 Local Packing Density

One of the simplest and most popular measures of local packing density in protein structure is the so-called site-specific Contact Number (CN) introduced and discussed by many authors in recent years (e.g., [62]) and frequently used in Elastic Network Models as measure of connectivity between individual sites in proteins. In its simplest mathematical form, the CN for a given site in protein is defined as the number of amino acids within a spherical neighborhood of fixed radius R_C centered at the site of interest [25]. Individual sites are generally represented by the coordinates of C_α backbone atoms for the calculation of CN. A major problem with the traditional definition of contact number however, is the existence of the arbitrary parameter R_C in the definition of CN. There is no consensus on the optimal value of this cutoff distance, although it is typically chosen in the range $7\mathring{A}$ to $13\mathring{A}$ [25, 64].

In an attempt to provide a more general definition of CN, some studies (e.g., [64]) have already suggested an alternative definition measure of local packing density known as the Weighted Contact Number (WCN): For a given site i in a protein of length N, WCN_i is defined as the sum of the inverse-squared of distances between the amino acid of interest and all other sites in protein,

$$WCN_i = \sum_{j \neq i}^{N} \frac{1}{{r_{ij}}^2}. \tag{2.14}$$

Evidently, the Weighted Contact Number definition encapsulates more information about the protein structure, including potential long-range interactions among distant amino acids, than the simple definition of CN. Indeed, it will be shown in the following chapters that WCN is virtually always a better predictor of other site-specific structural and evolutionary properties of proteins.

The local packing density measures have an intimate relationship with the local flexibility measures of proteins. Indeed, within the framework of Gaussian Network Models (GNM) [3] one can derive an inverse power-law relationship between packing density and flexibility of amino acid sites in proteins. Within the framework of GNM, the structure of a protein is represented by a set of nodes that are linked to each other via a set of springs of the same or varying constants. Each node represents a single site in the protein structure. In the simplest model, one can assume that each node is connected to any other node that are within the spherical neighborhood of radius R_C of each other, and the spring constants are all the same for all inter-node connections. In this case, the potential energy of a protein of N sites can be written as,

$$V = \frac{1}{2}\left(\sum_{i,j}^{N} \Gamma_{ij} \left[(\Delta X_i - \Delta X_j)^2 + (\Delta Y_i - \Delta Y_j)^2 + (\Delta Z_i - \Delta Z_j)^2 \right] \right) \tag{2.15}$$

in which the indices i & j refer to the i^{th} & j^{th} amino acids that interacting with each other, and $\Delta X, \Delta Y, \Delta Z$ represent the positional displacements of amino acids from their equilibrium positions in the three dimensional space. The symbol Γ_{ij} represents the ij^{th} element of the Kirchhoff connectivity matrix, which determines the interaction strength of the two amino acids. In the simplest scenario where all amino acids within a fixed spherical neighborhood have the same interaction strength (i.e., the same force constant γ), the Kirchhoff matrix can be written as,

$$\Gamma_{ij} = \begin{cases} -\gamma & \text{if } i \neq j,\ R_{ij} \lesssim R_C \\ 0 & \text{if } i \neq j,\ R_{ij} \gtrsim R_C \\ -\sum_{j,j\neq i} \Gamma_{ij} & \text{if } i = j \end{cases} \tag{2.16}$$

Since the GNM potential is isotropic and Gaussian, the fluctuation probability distributions of individual sites in all three spatial directions are independent and multiplicative. Taking the X-axis as an example, the probability of finding a protein of sequence length N in a state of fluctuation,

$$\Delta\mathbf{X} = \big[\Delta X_1, ..., \Delta X_N\big], \tag{2.17}$$

is given by,

$$p\big(\Delta\mathbf{X}\big) \propto \exp\bigg[-\frac{1}{2k_BT}\Big(\Delta\mathbf{X}^T\big(\mathbf{\Gamma}\big)\Delta\mathbf{X}\Big)\bigg], \tag{2.18}$$

in which k_B stands for the Boltzmann constant, and T is the absolute temperature. Therefore, the average fluctuations along the X-axis can be obtained

as,

$$\begin{aligned}\langle \Delta\mathbf{X}\Delta\mathbf{X}^T\rangle &= \int \Delta\mathbf{X}\Delta\mathbf{X}^T p\big(\Delta\mathbf{X}\big)\mathrm{d}\Delta\mathbf{X} \\ &= k_B T\big(\mathbf{\Gamma}^{-1}\big). \end{aligned} \tag{2.19}$$

Thus by symmetry, the total average fluctuations of a given site i in protein can be written as,

$$\langle \mathbf{\Delta R_i}^2\rangle = 3k_B T\big(\mathbf{\Gamma}^{-1}\big)_{ii}. \tag{2.20}$$

The diagonal element of the of Kirchhoff matrix is simply a measure of the local packing density as defnied in the previous paragraphs. In the case of weighted contact number as the maeasure of local packing density, the cutoff radius $R_C \to \infty$ and the the force constants become residue-residue dependent, that is, $\gamma \to \gamma_{ij}$. However, the final result in 2.20 remains intact.

The inverse of Kirchhoff matrix in 2.20 can be decomposed into the sum of a diagonal matrix (Γ_1) with the matrix of non-diagonal elements ($\mathbf{\Gamma}_2$), and therefore expanded to obtain,

$$\begin{aligned}\mathbf{\Gamma}^{-1} &= \Big[\mathbf{\Gamma}_1 + \mathbf{\Gamma}_2\Big]^{-1} = \Big[\mathbf{\Gamma}_1\Big(\mathbf{E} + \mathbf{\Gamma}_1^{-1}\mathbf{\Gamma}_2\Big)\Big]^{-1} \\ &= \mathbf{\Gamma}_1^{-1} - \mathbf{\Gamma}_1^{-1}\mathbf{\Gamma}_2\mathbf{\Gamma}_1^{-1} + ... \\ &\sim \mathbf{\Gamma}_1^{-1} \end{aligned} \tag{2.21}$$

under the assumption that the terms involving $\mathbf{\Gamma}_1^{-1}\mathbf{\Gamma}_2$ are small compared to the identity matrix $\mathbf{E}$. Thus, Equation 2.20 in combination with Equation 2.21 simply imply that the fluctuations of individual sites in proteins are inversely proportional to the local packing density around the sites. Figure 2.5 illustrates that there is indeed such an inverse relation between the Weighted Contact Number of individual sites proteins as a measure of local packing density, and the average B factor of sites as a measure of site fluctuations.

2.7 Amino Acid Hydrogen Bond Strength

Hydrogen bond (H-bond) is a strong type of dipole-dipole electrostatic attraction between the Hydrogen atom (N–H) in the backbone of one amino acid and the strongly electronegative Oxygen atom (C–O) in the backbone of another amino acid. The interaction energy in units of $kcal/mol$ can be approximated by assuming partial charges $+q_1, -q_1$ on C–O atoms and $-q_2, +q_2$ on N–H atoms using the following approximate relation,

$$E = f \times q_1 \times q_2 \left(\frac{1}{r_{\mathrm{ON}}} + \frac{1}{r_{\mathrm{CH}}} - \frac{1}{r_{\mathrm{OH}}} - \frac{1}{r_{\mathrm{CN}}} \right), \tag{2.22}$$

in which r_{AB} stands for the distance between atoms A & B in units of Angstroms, $q_1 = 0.42q_e$ $q_1 = 0.42q_e$ with q_e standing for the charge of electron, and $f = 332$ serving as a dimensionless normalizing factor. A rather strong H-bond corresponds to an energy $E \lesssim -3$kcal/mol [48].

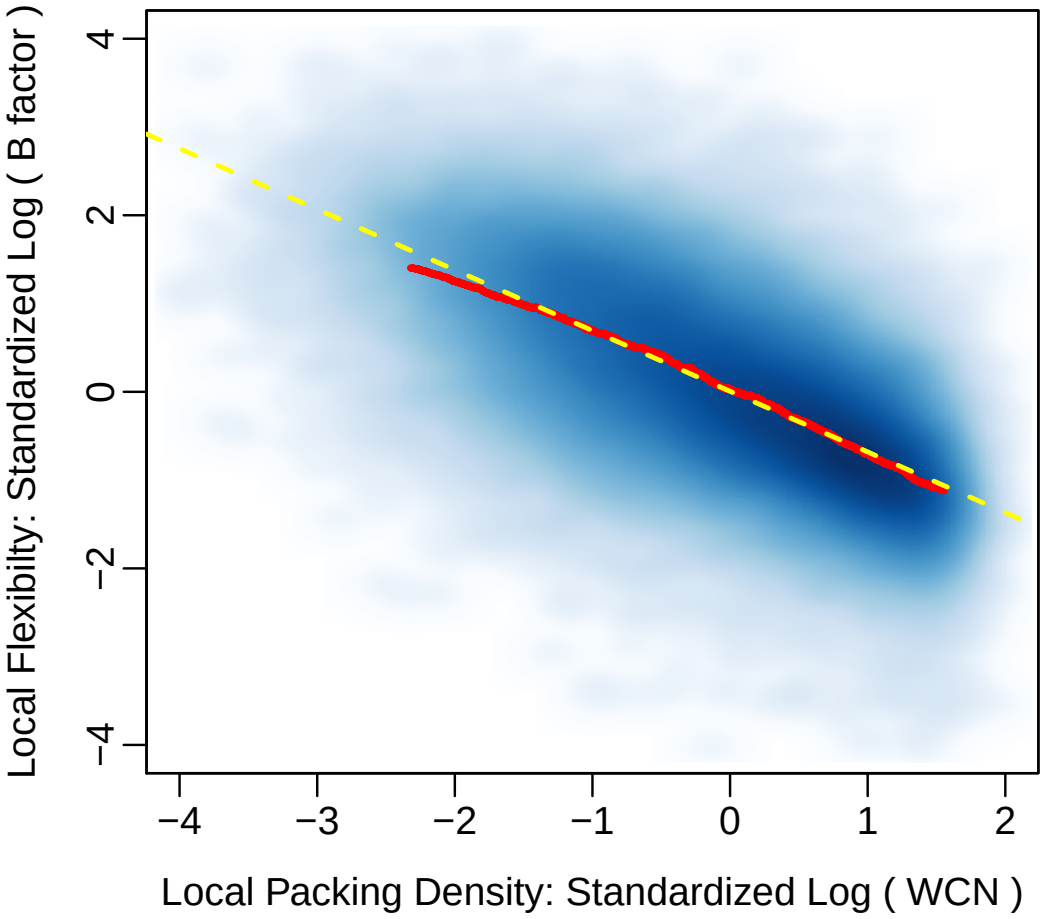

Figure 2.5: An illustration of the linear relationship between the standardized logarithm of local packing density as measured by the Weighted Contact Number (WCN) and the standardized logarithm of local flexibility of individual sites in proteins, as measured by the average side-chain B factor. The red curve is the result of adjacent averaging of 77150 amino acid sites in 209 enzymatic proteins (c.f., Chapter 4), every 3000 sites. The blue shaded area is 2D heat map of the scatter of 77150 sites about the average red curve. The dashed yellow line a symmetric Deming linear regression fit to the average curve, with a slope of $m \sim -0.7$. The observed deviation in the value of the slope from -1 is potentially the result of several contributing factors, most importantly, the systematic biases and errors in B factors as measures of local flexibility and the approximations made in obtaining the relationship between packing density and flexibility (Equations 2.20 & 2.21).

Chapter 3

Structural Determinants of Sequence Evolution in Viral Proteins: Buriedness, Packing, Flexibility, and Design

3.1 Introduction

Patterns of amino-acid sequence variation in protein-coding genes are shaped by the structure and function of the expressed proteins [63, 71, 122]. As the most basic reflection of this relationship, buried residues in proteins tend to be more evolutionarily conserved than exposed residues [14, 31, 78, 84]. More specifically, when evolutionary variation is plotted as a function of Relative Solvent Accessibility (RSA, a measure of residue buriedness), the relationship falls, on average, onto a straight line with a positive slope [23, 24, 89, 101]. Importantly, however, this relationship represents an average over many sites and many proteins. At the level of individual sites in individual proteins, RSA is often only weakly correlated with evolutionary variation [73, 74, 131].

Other structural measures, such as residue contact number (CN), have also been shown to correlate with sequence variability [23, 61, 131], and some have argued that CN predicts evolutionary variation better than RSA [130, 131]. Because CN may be a proxy for residue and site-specific backbone flexibility [37], a positive trend between local structural variability and sequence

variability may also exist [131]. Indeed, several authors have suggested that such protein dynamics may play a role in sequence variability [65,71,81]. However, a recent paper argued against the flexibility model, on the grounds that evolutionary rate is not linearly related to flexibility [42].

While RSA and CN can be calculated in a straightforward manner from individual crystal structures, measures of structural flexibility, either at the side-chain or the backbone level, are more difficult to obtain. Two viable approaches to measuring structural flexibility are (i) examining existing structural data or (ii) simulating protein dynamics. NMR ensembles may approximate physiologically relevant structural fluctuations. Similar fluctuations are observed in ensembles of homologous crystal structures [18,68]. The thermal motion of atoms in a crystal are recorded in B factors, which are available for every atom in every crystal structure. To measure protein fluctuations using a simulation approach, one can either use coarse-grained modeling, e.g. via Elastic Network Models [100], or atom-level modeling, e.g. via molecular dynamics (MD) [50]. However, it is not well understood which, if any, of these measures of structural flexibility provide insight into the evolutionary process, in particular into residue-specific evolutionary variation.

Here, I provide a comprehensive analysis of the extent to which numerous different structural quantities predict evolutionary sequence (amino-acid) variation. I consider two measures of evolutionary sequence variation: site entropy, as calculated from homologous protein alignments, and evolutionary rate. As structural predictors, I included buriedness (RSA), packing

density (CN), and measures of structural flexibility, including B factors, several measures of backbone and side-chain variability obtained from MD simulations, and backbone variability obtained from alignments of homologous crystal structures. I additionally consider site variability, as predicted from computational protein design with Rosetta.

On a set of nine viral proteins, RSA and CN generally performed better at predicting evolutionary site variation than did either measures of structural flexibility or computational protein design. Among the measures of structural flexibility, measures of side-chain variability performed better than do measures of backbone variability, possibly because the former are more tightly correlated with residue packing. Finally, site variability predicted from computational protein design performed worse than the best-performing measures of structural fluctuations.

3.2 Materials and Methods

3.2.1 Sequence Preparation, Alignments, and the Calculation of Evolutionary Rates

All viral sequences except influenza sequences were retrieved from `http://hfv.lanl.gov/components/sequence/HCV/search/searchi.html`. The sequences were truncated to the desired genomic region but not in any other way restricted. Influenza sequences were downloaded from `http://www.fludb.org/brc/home.spg?decorator=influenza`. Only human influenza A, H1N1, were considered in this work, excluding H1N1 sequences derived from the 2009

Swine Flu outbreak or any sequence from before 1998, but with no geographic restrictions.

For all viral sequences, any sequence that was not in reading frame were removed, in addition to any sequence which was shorter than 80% of the longest sequence for a given viral protein (so as to remove all partial sequences), and any sequence containing any ambiguous characters. Alignments were constructed using amino-acid sequences with MAFFT [51,52], specifying the `--auto` flag to select the optimal algorithm for the given data set, and then back-translated to a codon alignment using the original nucleotide sequence data.

To assess site-specific sequence variability in amino-acid alignments, the Shannon entropy (H_i) at each alignment column i is calculated according to Eqn. 2.1. For each alignment, the evolutionary rates are also calculated, as described in [112]. In brief, a phylogeny for each codon alignment in RAxML [113] is generated using the GTRGAMMA model. Then, using the codon alignment and phylogeny, the site-specific evolutionary rates are inferred with a Random Effects Likelihood (REL) model, using the HyPhy software [55]. The REL model was a variant of the GY94 evolutionary model [32] with five ω rate categories as free parameters. An Empirical Bayes approach [128] is employed to infer ω values for each position in the alignment. These ω values represent the evolutionary-rate ratio dN/dS at each site.

3.2.2 Protein Crystal Structures

A total of 9 viral protein structures were selected for analysis, as tabulated in Table 3.1. Sites in the PDB structures were mapped to sites in the viral sequence alignments via a custom-built python script that creates a consensus map between a PDB sequence and all sequences in an alignment.

For each of the viral proteins, homologous structures were identified using the `blast.pdb` function of the R package Bio3D [34]. BLAST hits were retained if they had $\geq 35\%$ sequence identity and $\geq 90\%$ alignment length. Among the retained hits, sets of homologous structures were subsequently identified with unique sequences and with mutual pairwise sequence divergences of $\geq 2\%$, $\geq 5\%$, and $\geq 10\%$.

Table 3.1: PDB structures considered in this study.

Viral Protein	PDB ID	Chain	Sequence Length	Number of Sequences
Hemagglutinin Precursor	1RD8	AB	503	1039
Dengue Protease Helicase	2JLY	A	451	2362
West Nile Protease	2FP7	B	147	237
Japanese Encephalitis Helicase	2Z83	A	426	145
Hepatitis C Protease	3GOL	A	557	1021
Rift Valley Fever Nucleoprotein	3LYF	A	244	95
Crimean Congo Nucleocapsid	4AQF	B	474	69
Marburg RNA Binding Domain	4GHA	A	122	42
Influenza Nucleoprotein	4IRY	A	404	943

3.2.3 Molecular Dynamics Simulations

Molecular dynamics (MD) simulations were carried out using the GPU implementation of the *Amber12* simulation package [99] with the most recent release of the Amber fixed-charge force field (ff12SB; c.f., AmberTools13 Manual). Prior to MD production runs, all PDB structures were first solvated in a box of TIP3P water molecules [47] such that the structures were at least $10\mathring{A}$ away from the box walls. Each individual system was then energy minimized using the steepest descent method for 1000 steps, followed by conjugate gradient for another 1000 steps. Then, the structures were constantly heated from 0K to 300K for 0.1ns, followed by 0.1ns constant pressure simulations with positional harmonic restraints on all atoms to avoid instabilities during the equilibration process. The systems were then equilibrated for another 5ns without positional restraints, each followed by 15ns of production simulations for subsequent post-processing and analyses. All equilibration and production simulations were run using the SHAKE algorithm [97]. Langevin dynamics were used for temperature control.

3.2.4 Measures of Buriedness, Packing Density, and Structural Flexibility

As a measure of residue buriedness, the Relative Solvent Accessibility (RSA) is used. To calculate RSA, first the Accessible Surface Area (ASA) for each residue in each protein is calculated using the DSSP software [49]. The ASA values are then normalized by the theoretical maximum ASA of

each residue [117] to obtain RSA. Two measures of local packing density were considered, contact number (CN) and weighted contact number (WCN). I calculated CN for each residue as the total number of Cα atoms surrounding the Cα atom of the focal residue within a spherical neighborhood of a predefined radius r_0. Following Yeh et al. (2014) [131], I used $r_0 = 13\mathring{A}$. I calculated WCN (Eqn. 2.14)as the total number of surrounding Cα atoms for each focal residue, weighted by the inverse square separation between the Cα atoms of the focal residue and the contacting residue, respectively [106].

In most analyses, the inverse of CN and/or WCN were actually used, iCN = 1/CN and iWCN = 1/WCN. Note that for Spearman correlations, which is used throughout the entire work here, replacing a variable by its inverse changes the sign of the correlation coefficient but not the magnitude.

As measures of structural flexibility, the Root-Mean-Square Fluctuations (RMSF) of C_α backbone atoms in amino acids (Eqn. 2.11) were measured from Molecular Dynamics simulations, variability in backbone and side-chain dihedral angles from MD simulations, and B factors from PDB files were considered. To calculate RMSF from homologous structures, the structures were first aligned using the Bio3D package [34], based upon which RMSF is calculated according to Eqn. 2.12, using Cα atoms as the representations of individual sites in proteins and site-specific weights (w_i) to correct for potential phylogenetic relationship among the aligned structures. The weights w_i were calculated using BranchManager [115], based on phylogenies built with RAxML as before.

To assess variability in backbone and side-chain dihedral angles, the quantities Var(ϕ), Var(ψ), and Var(χ_1) were used. The variance of a dihedral angle was defined according to the most common definition in directional statistics as given by Eqn. 2.13

B factors were extracted from the crystal structures. Only B factors of the Cα atoms of amino acids in protein were considered.

3.2.5 Sequence Entropy from Designed Proteins

Designed entropy was calculated as described [45]. In brief, proteins were designed using RosettaDesign (Version 39284) [59] using a flexible backbone approach. This was done for all PDB structures in Table 3.1 as initial template structures. For each template, a backbone ensemble was created using the Backrub method [110]. The temperature parameter in Backrub was set to 0.6, allowing for an intermediate amount of flexibility. It has been previously found in a different data set that intermediate flexibility gives the highest congruence between designed and observed site variability [45]. For each of the 9 template structures, 500 proteins were designed.

All details of simulations, input/output files, and scripts for subsequent analyses are available to view or download at `https://github.com/clauswilke/structural\_prediction\_of\_ER`.

3.3 Results

3.3.1 Dataset and Structural Variables Considered

Our goal in this work was to determine which structural properties best predict amino-acid variability at individual sites in viral proteins. To this end, I selected 9 viral proteins for which I had both high-quality crystal structures and abundant sequences to assess evolutionary sequence variation (Table 3.1). I quantified evolutionary variability in two ways: by calculating sequence entropies for each alignment column, and by calculating site-specific evolutionary-rate ratios $\omega = dN/dS$ (see Methods for details). Throughout this paper, I primarily report results obtained for sequence entropy. Results for ω were largely comparable, with some specific caveats detailed below.

As predictors of evolutionary variability, I considered buriedness, packing density, and residue flexibility. I additionally considered the variation seen in computationally designed protein variants. Buriedness quantifies the extent to which a residue is protected from solvent. I determined residue buriedness by calculating the relative solvent accessibility (RSA), which represents the relative proportion of a residue's surface in contact with solvent.

Packing density quantifies how many other residues a given residue interacts with. I determined packing density by calculating contact number (CN) and weighted contact number (WCN). CN counts the number of contacts within a sphere of a given radius around the α-carbon of the focal residue, while WCN weights contacts by the distance between the two residues. Residue buriedness and packing density tend to be (anti-)correlated but mea-

sure qualitatively different properties of a residue. In particular, in the core of a protein, buriedness is always zero but packing density can vary. Because contact numbers decline as relative solvent accessibility increases, I replaced CN and WCN with their inverses, iCN = 1/CN and iWCN = 1/WCN, in most analyses. Importantly, as Spearman rank correlations were used, this substitution only changed the sign of correlations but not the magnitude.

Measures of structural flexibility assess the extent to which a residue fluctuates in space as a protein undergoes thermodynamic fluctuations in solution. I quantified these fluctuations using several different measures. I considered B factors, which measure the spatial localization of individual atoms in a protein crystal, RMSF, the root mean-square fluctuation of the Cα atom over time, and variability in side-chain and backbone dihedral angles, including Var(χ_1), Var(ϕ), and Var(ψ). I employed two broad approaches, one using PDB crystal structures and one using molecular dynamics (MD) simulations, to obtain these measurements. Crystal structures yielded measures for B factors and RMSF; I obtained B factors from individual protein crystal structures, given in Table 3.1, and I calculated RMSF from aligned homologous crystal structures for those proteins which had sufficient sequence variation among crystal structures (see Methods and Table 3.2 for details). MD simulations yielded measures for RMSF and variability in residue dihedral and side-chain angles. More specifically, I simulated MD trajectories for all crystal structures in Table 3.1. For each protein, I equilibrated the structure, simulated 15ns of chemical time, and recorded snapshots of the simulated structure every 10ps

(see Methods for details). I obtained RMSF and angle variabilities from these snapshots. Additionally, I calculated time-averaged values of RSA, CN, and WCN. I also refer to these time-averaged measures as MD RSA, MD CN, and MD WCN, respectively. Unless specified otherwise, all results reported below were obtained using MD RSA, MD CN, and MD WCN.

Table 3.2: Availability of homologous crystal structures. Although most viral proteins have many PDB structures available, the sequence divergence among these structures is low. Therefore, when calculating RMSF from crystal structures, I considered only those proteins with at least five homologous structures at 5% pairwise sequence divergence (highlighted in bold).

Viral Protein	BLAST hits[a]	Unique sequences			
		all	$\geq 2\%$[b]	$\geq 5\%$[b]	$\geq 10\%$[b]
Hemagglutinin Precursor	63	17	10	**9**	7
Dengue Protease Helicase	31	13	7	**7**	7
West Nile Protease	21	16	10	**7**	6
Japanese Encephalitis Helicase	31	12	7	**7**	7
Hepatitis C Protease	302	33	10	**5**	4
Rift Valley Fever Nucleoprotein	95	9	5	**5**	5
Crimean Congo Nucleocapsid	7	4	3	2	2
Marburg RNA Binding Domain	63	9	5	3	3
Influenza Nucleoprotein	69	15	4	4	2

[a] BLAST hits against all sequences in the PDB, excluding hits with $< 35\%$ sequence identity and $< 90\%$ alignment length

[b] Unique sequences at indicated minimum pairwise sequence divergence

As an alternative to predicting evolutionary variation from simple structural measures such as contact density or backbone flexibility, one can also predict evolutionary variation via a protein-design approach [16, 45, 82]. In this case, one takes the protein structure of interest, replaces all residue side chains with randomly-chosen alternatives, and uses a coarse-grained or atom-level en-

ergy function to assess which side-chain choices are consistent with the backbone conformation of the focal structure. I have recently used this approach to compare natural and designed sequence variability in cellular proteins [45], and I have found that (i) flexible-backbone design, where small backbone movements are allowed during the design phase, outperformed fixed-backbone design, and (ii) intermediate backbone flexibility, obtained via an intermediate design temperature, produced the highest congruence between designed and natural sequences. Similarly, [16] had previously found that an intermediate temperature parameter gave the best agreement between designed and natural sequences in their model. Inspired by these prior results, I investigated here how protein design performed relative to simpler structural quantities. For all proteins in our study (Table 3.1), I used the Rosetta protein-design platform [59] to generate 500 designed variants. I then calculated the sequence entropy at each alignment position of the designed variants. I refer to the resulting quantity as the *designed entropy*. I chose a design temperature of $T = 0.6$, which was near the optimal range in our previous work [45].

3.3.2 Evaluating Structural Predictors of Sequence Evolution

I began by comparing the Spearman correlations of sequence entropy with six different measures of local structural flexibility: B factors, RMSF obtained from MD simulations (MD RMSF), and RMSF obtained from crystal structures (CS RMSF), and variability in backbone and side-chain dihedral angles (ϕ, ψ, and χ_1). The correlation strengths of these quantities with

entropy are shown in Figure 3.1. Significant correlations ($P < 0.05$) are shown with filled symbols, and non-significant correlations are shown with empty symbols ($P \geq 0.05$). I found that the variability in backbone dihedral angles, Var(ϕ) and Var(ψ), explained the least variation in sequence entropy, while the variability in the side-chain dihedral angle, Var(χ_1), explained, on average, more variation in sequence entropy than did any other measure of structural flexibility. B factors and the two measures of RMSF explained, on average, approximately the same amount of variation in entropy, even though the results for individual proteins were somewhat discordant (see also next sub-section).

Based on results from the above analysis, I proceeded to compare the relative explanatory power among the best-performing measures of structural flexibility (Var(χ_1), MD RMSF, and B factors) with buriedness (RSA), packing density (iWCN), and designed entropy. Figure 3.2 shows the Spearman correlation coefficients between sequence entropy and each of the aforementioned quantities, for all proteins in our analysis. In this figure, several patterns emerged. First, nearly all correlations were positive and most were statistically significant, with the main exception of the Marburg virus RNA binding domain (PDB ID 4GHA). This protein only showed a single significant negative correlation between sequence entropy and Var(χ_1). Second, correlations were generally weak, such that no correlation coefficient exceeded 0.4. Third, on average, correlations were strongest for RSA and iWCN, yielding average correlations of $\rho = 0.23$ and $\rho = 0.22$, respectively. Fourth, designed entropy performed worse than RSA or iWCN as a predictor of evolutionary sequence

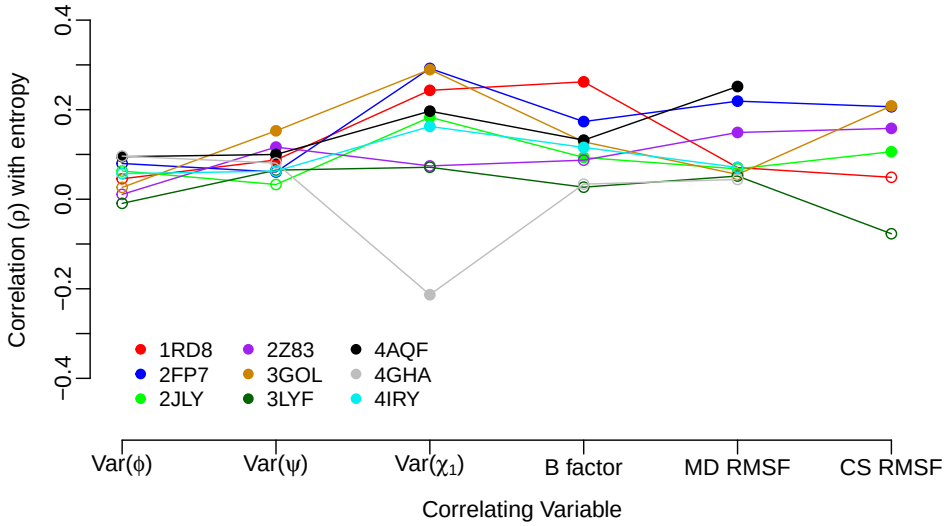

Figure 3.1: Spearman correlation of sequence entropy with measures of structural variability. Each symbol represents one correlation coefficient for one protein structure. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. The quantities Var(ψ), Var(ϕ), Var(χ_1), and MD RMSF were obtained as time-averages over 15ns of MD simulations. B factors were obtained from individual crystal structures. CS RMSF values were obtained from alignments of homologous crystal structures when available. Almost all structural measures of variability correlate weakly, but significantly, with sequence entropy.

variability, but it performed roughly the same as the three flexibility measures in this figure; the values of designed entropy, Var(χ_1), MD RMSF, and B factors showed average correlations of $\rho = 0.13$, $\rho = 0.14$, $\rho = 0.11$, and $\rho = 0.12$, respectively.

3.3.3 MD Time-Averages vs. Crystal-Structure Snapshots

Except for analyses involving B factors and CS RMSF, I obtained structural measures by averaging quantities over MD trajectories. This approach, however, did not reflect conventional practice for measuring RSA, CN, or WCN, which are typically measured from individual crystal structures. Therefore, I examined whether MD time-averages differed in any meaningful way from estimates obtained from crystal structures, and whether these estimates differed in their predictive power for evolutionary sequence variation.

As shown in Table 3.3, RSA, CN, and WCN from crystal structures were highly correlated with their corresponding MD trajectory time-averages, for all protein structures I examined (Spearman correlation coefficients of > 0.9 in all cases). Furthermore, the correlation coefficients I obtained when comparing the crystal-structure based measures to sequence entropy were virtually identical to coefficients obtained from the MD trajectory correlations (Figure 3.3A-C). Thus, in terms of predicting evolutionary variation, RSA, CN, and WCN values obtained from static structures performed as well as their MD equivalents averaged over short time scales.

By contrast, correlations between corresponding MD RMSF to CS

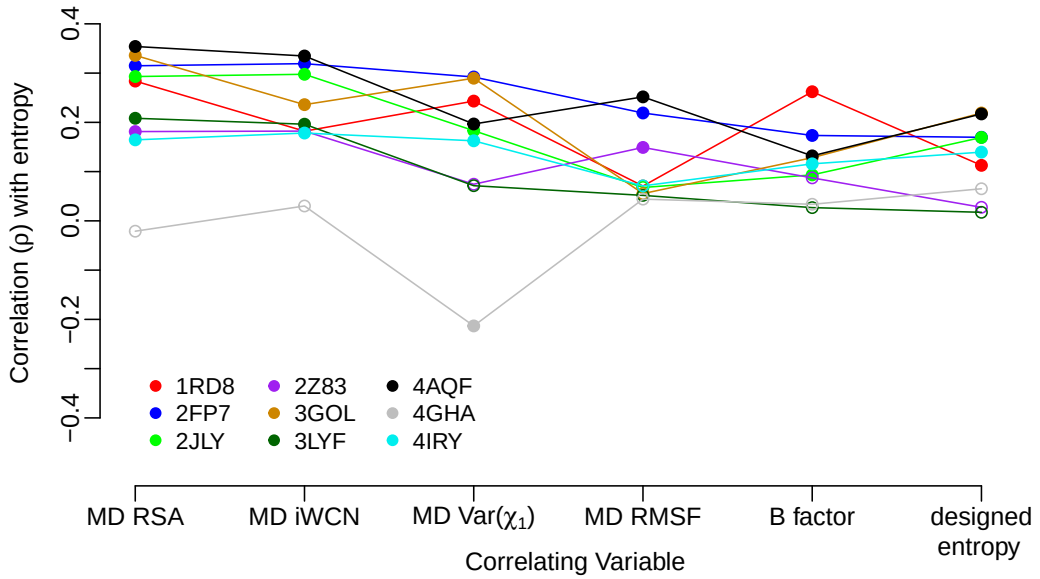

Figure 3.2: Spearman correlation of sequence entropy with measures of buriedness, packing density, and structural flexibility, as well as with designed entropy. Each symbol represents one correlation coefficient for one protein structure. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. The quantities MD RSA, MD iWCN, MD Var(χ_1), and MD RMSF were calculated as time-averages over 15ns of MD simulations. B factors were obtained from crystal structures, and designed entropy was obtained from protein design in Rosetta. Compared to the measures of structural variability and to designed entropy, MD RSA and MD iWCN consistently show stronger correlations with sequence entropy. Note that results for MD iWCN are largely identical to those for MD iCN, so only MD iWCN was included here.

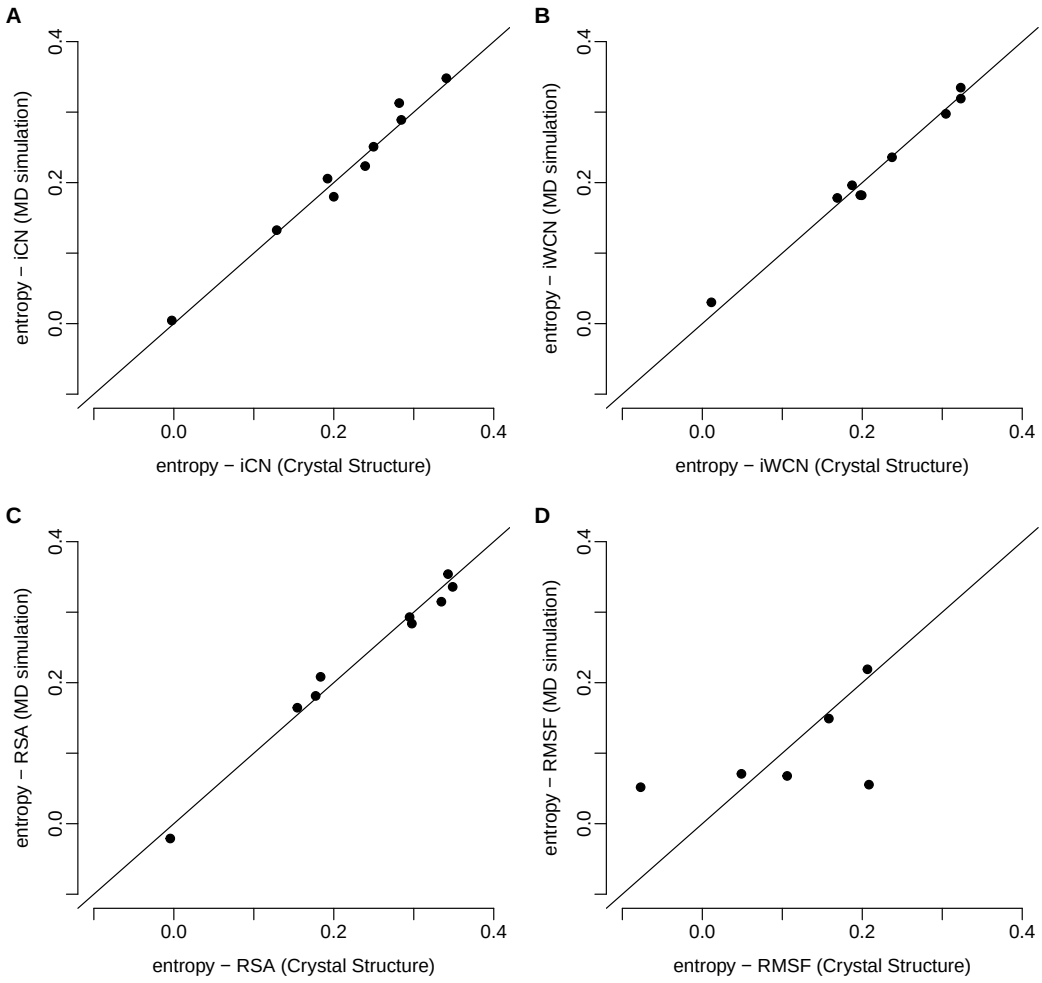

Figure 3.3: Spearman correlations of sequence entropy with MD-derived and crystal-structure derived structural measures. The vertical axes in all plots represent the Spearman correlation of sequence entropy with one structural variable obtained from $15ns$ of molecular dynamics (MD) simulations. The horizontal axes represent the Spearman's rank correlation coefficient of sequence entropy with the same structural variable as in the vertical axes but measured from protein crystal structures. Each dot represents one correlation coefficient for one protein structure. The quantities iCN, iWCN, and RSA have nearly identical predictive power for sequence entropy regardless of whether they are derived from MD simulations or from crystal structures. By contrast, MD RMSF yielded very different correlations than did CS RMSF.

Table 3.3: Correlations between quantities obtained from MD trajectories and from crystal structures. For each quantity and each protein, I calculated the Spearman correlation ρ between the values obtained from MD time averages and the values obtained from viral protein crystal structures. Note that crystal structures for all nine proteins were used for RSA, CN, and WCN calculations, but only the six proteins for which I had sufficient crystal structure variability were used for CS RMSF. I then calculated the minimum, maximum, mean, and standard deviation of these correlations.

Quantity	min ρ	max ρ	$\langle\rho\rangle$	SD(ρ)
RSA	0.937	0.981	0.948	0.012
CN	0.964	0.993	0.976	0.008
WCN	0.973	0.991	0.984	0.006
RMSF	0.218	0.723	0.502	0.181

RMSF measures were sometimes quite different, with correlation coefficients ranging from 0.218 to 0.723 (Table 3.3). Consequently, for the two proteins for which MD RMSF was the least correlated with CS RMSF (hepatitis C protease and Rift Valley fever nucleoprotein), the strength of correlation between site entropy and RMSF depended substantially on how RMSF was calculated (Figures 3.1 and 3.3D).

Finally, I examined whether correlations between sequence entropy and B factors or the two RMSF measures were comparable (Figure 3.4). Again, I found that correlations between sequence entropy and B factors were generally different from those obtained for both MD RMSF and CS RMSF. This result highlighted that, while B factors, MD RSMF, and CS RMSF all measure backbone flexibility, they each contain distinct information about evolutionary sequence variability in our data set.

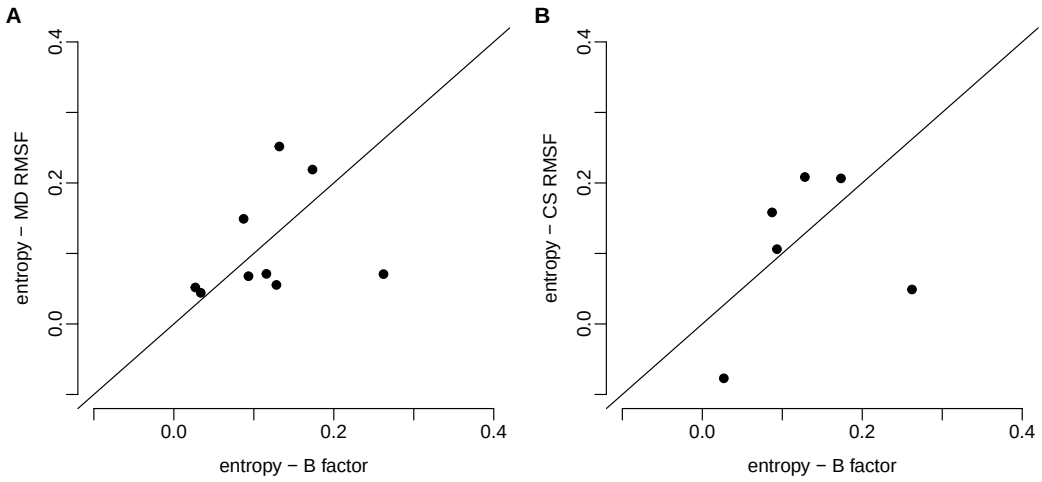

Figure 3.4: Spearman correlations of sequence entropy with measures of structural variability. Vertical and horizontal axes represent Spearman correlations of the indicated quantities. Each dot represents one correlation coefficient for one protein structure. MD RMSF, CS RMSF, and B factors all explain different amounts of variance in sequence entropy for different proteins.

3.3.4 Sequence Entropy vs. Evolutionary-Rate Ratio ω

In the previous subsections, I used sequence entropy as a measure of site-wise evolutionary variation. While sequence entropy is a simple and straightforward measure of site variability, it has two potential drawbacks. First, while measured from homologous protein alignments, sequence entropy doesn't correct for the phylogenetic relationship of those alignment sequences. Hence, entropy can be biased if some parts of the phylogeny are more densely sampled than others. Second, entropy does not take the actual substitution process into account. As a result, a single substitution near the root of the tree can result in a comparable entropy to a sequence of substitutions toggling back and forth between two amino acids.

To consider an alternative quantity of evolutionary variation that doesn't suffer from either of these drawbacks, I calculated the evolutionary-rate ratio $\omega = dN/dS$ for all proteins at all sites, and repeated all analyses with ω instead of entropy. I found that results generally carried over, but with somewhat weaker correlations. Figure 3.5 plots, for each protein, the Spearman correlations between ω and our various predictors versus the correlation between entropy and our predictors. Most data points fall below the $x = y$ line and are shifted downwards by approximately 0.1. Thus, correlations of structural quantities and designed entropy with ω are, on average, approximately 0.1 smaller than correlations of the same quantities with sequence entropy.

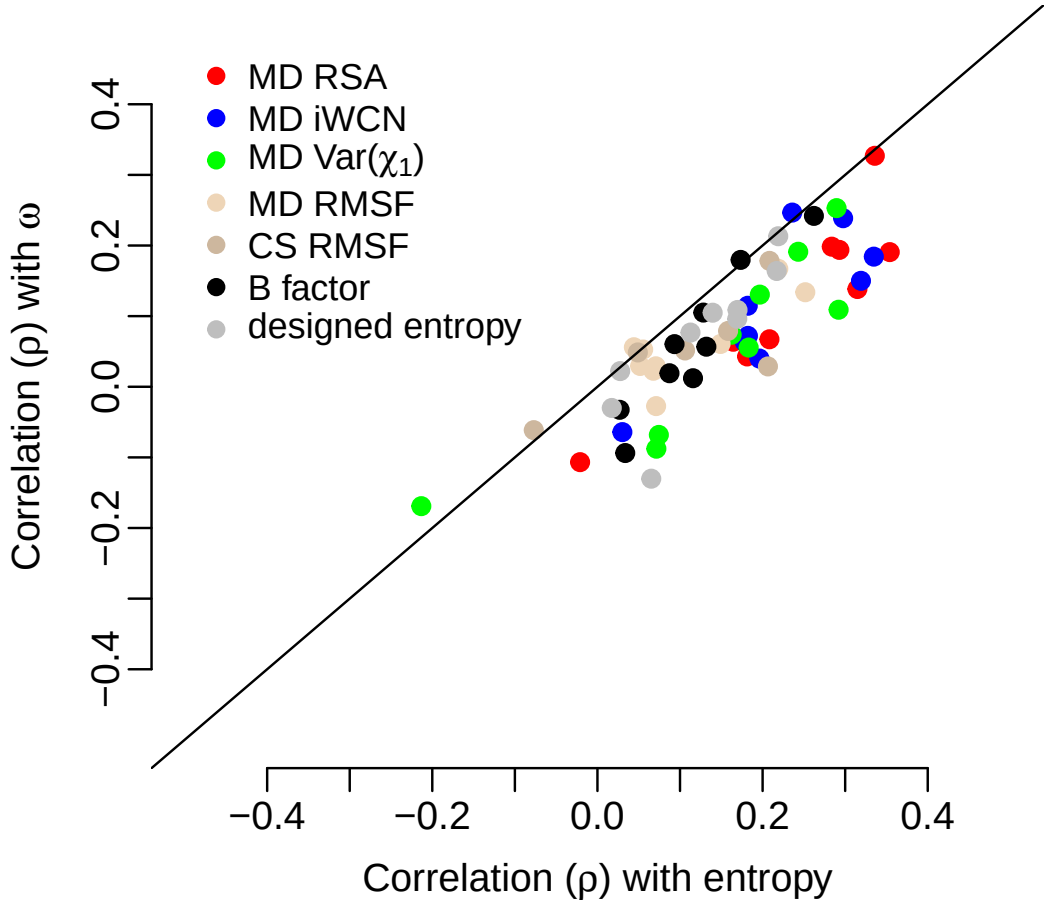

Figure 3.5: Spearman correlations of structural quantities with sequence entropy and with the evolutionary rate ratio ω. Nearly all points fall below the $x = y$ line, indicating that structural quantities generally predict as much as or more variation in sequence entropy than in ω.

3.3.5 Multivariate Analysis of Structural Predictors

The various structural quantities I have considered are by no means independent of each other. Measures of buriedness and packing density co-vary with each other, as do measures of structural flexibility. Further, the latter co-vary with the former, as does designed entropy. Therefore, I conducted a joint multivariate analysis, which included most structural quantities considered in this work. I employed this strategy to determine the extent to which these quantities contained independent information about sequence variability while additionally assessing whether combining multiple structural quantities yielded improved predictive power. I employed a principal component (PC) regression approach, which has previously been used successfully to disentangle genomic predictors of whole-protein evolutionary rates [5, 17]. For each analysis described below, I first carried out a PC analysis of the predictor variables (i.e., the structural quantities such as RSA and RMSF), and I subsequently regressed the response (either sequence entropy or ω) against the individual components. Note that variables were not rank-transformed for this analysis.

For a first PC analysis, I pooled all structural quantities and then regressed entropy against each PC separately, for all proteins in our data set. This strategy allowed us to analyze all proteins in our data set individually but in such a way that results were comparable from one protein to the next. I excluded CS RMSF from this analysis, so that I could include results from all nine viral proteins. The results of this analysis are shown in Figure 3.6. The first component (PC1) explained, on average, the largest amount of vari-

ation in sequence entropy (see Figure 3.6A). PC3 yielded the second-highest r^2 value, on average, while all other components explained very little variation in sequence entropy. When looking at the composition of the components, I found that RSA, iWCN, RMSF, and Var(χ_1) all loaded strongly on PC1, while PC2 and PC3 where primarily represented by designed entropy and B factors (see Figure 3.6B and C). RMSF also had moderate loadings on PC3. Interestingly, designed entropy and B factors load with equal signs on PC2 but with opposite signs on PC3.

I interpreted PC1 to represent a buriedness/packing-density component. By definition, PC1 measures the largest amount of variation among the structural quantities, and all structural quantities reflect to some extent the buriedness of residues and the number of residue-residue contacts. PC2 and PC3 were more difficult to interpret. Since designed entropy and B factors loaded strongly on both but with two different combinations of signs, I concluded that the most parsimonious interpretation was to consider PC2 as a component representing sites with high designed entropy and high spatial fluctuations (as measured by B factors) and PC3 representing sites with high designed entropy and low spatial fluctuations. Using these interpretations, our PC regression analysis suggested that of all the structural quantities considered here, residue buriedness/packing was the best predictor of evolutionary variation. Designed entropy was a useful predictor as well, but it tended to perform better at sites with low spatial fluctuations.

For a second PC analysis, I included the predictor CS RMSF, which

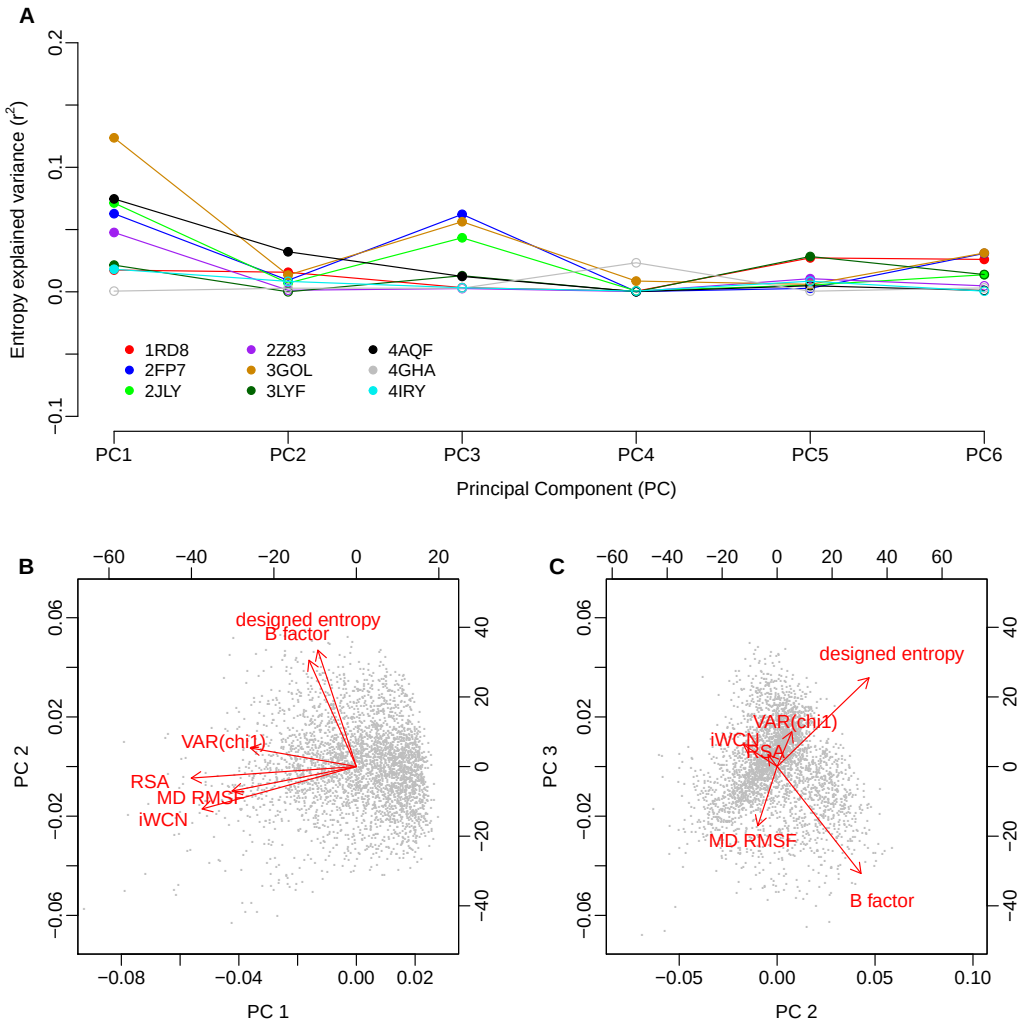

Figure 3.6: Principal Component (PC) Regression of sequence entropy against structural variables. **(A)** Variance in entropy explained by each principal component. For most proteins, PC1 and PC3 show the strongest correlations with sequence entropy. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. **(B)** and **(C)** Composition of the three leading components. Red arrows represent the loadings of each of the structural variables on the principal components; black dots represent the amino acid sites in the PC coordinate system. The variables RSA, iWCN, MD RMSF, and Var(χ_1) load strongly on PC1 and weakly on PC2, while B factor and designed entropy load strongly on PC2 and weakly on PC1.

therefore restricted the data set to include only six proteins (see Table 3.2). This analysis, which retained sequence entropy as the response variable, yielded comparable results to the first PC analysis. The main differences occurred in PC2 and PC3, where CS RMSF generally loaded in the opposite direction of B factor, and either in the same (PC2) or the opposite (PC3) direction of designed entropy (Figure 3.7).

Finally, I redid the two PC analyses described above, but instead with ω as the response variable (Figures 3.8 and 3.9). Again, these results were largely comparable to results from PC analyses with sequence entropy as the response.

3.4 Discussion

I have carried out a comprehensive analysis of the extent to which different structural quantities predict sequence evolutionary variation in nine viral proteins. I found that measures of buriedness and local packing generally performed better than did measures of structural flexibility. Further, the former measures also performed better than a computational protein-design approach that employed a sophisticated all-atom force field to determine allowed amino-acid distributions at each site. Finally, there was no difference in predictive power between structural quantities obtained from averaging structural quantities over 15ns of MD simulations versus taking the same quantities from individual crystal structures.

Our results are broadly in agreement with recent work by Echave and collaborators [42,131]. These authors found that RSA and CN showed comparable correlation strengths with evolutionary sequence variation [131]. Further, they demonstrated that the observed relationship between evolutionary variation and residue–residue contacts was not consistent with a flexibility model that puts evolutionary variability in proportion to structural flexibility [42]. Instead, a mechanistic stress model, in which amino-acid substitutions cause physical stress in proportion to the number of residue–residue contacts affected, could explain all the observed data [42].

The correlation strengths I observed were consistently lower than those observed previously [45, 131]. I believe that this result was due to our choice of analyzing viral proteins instead of the cellular proteins or enzymes used

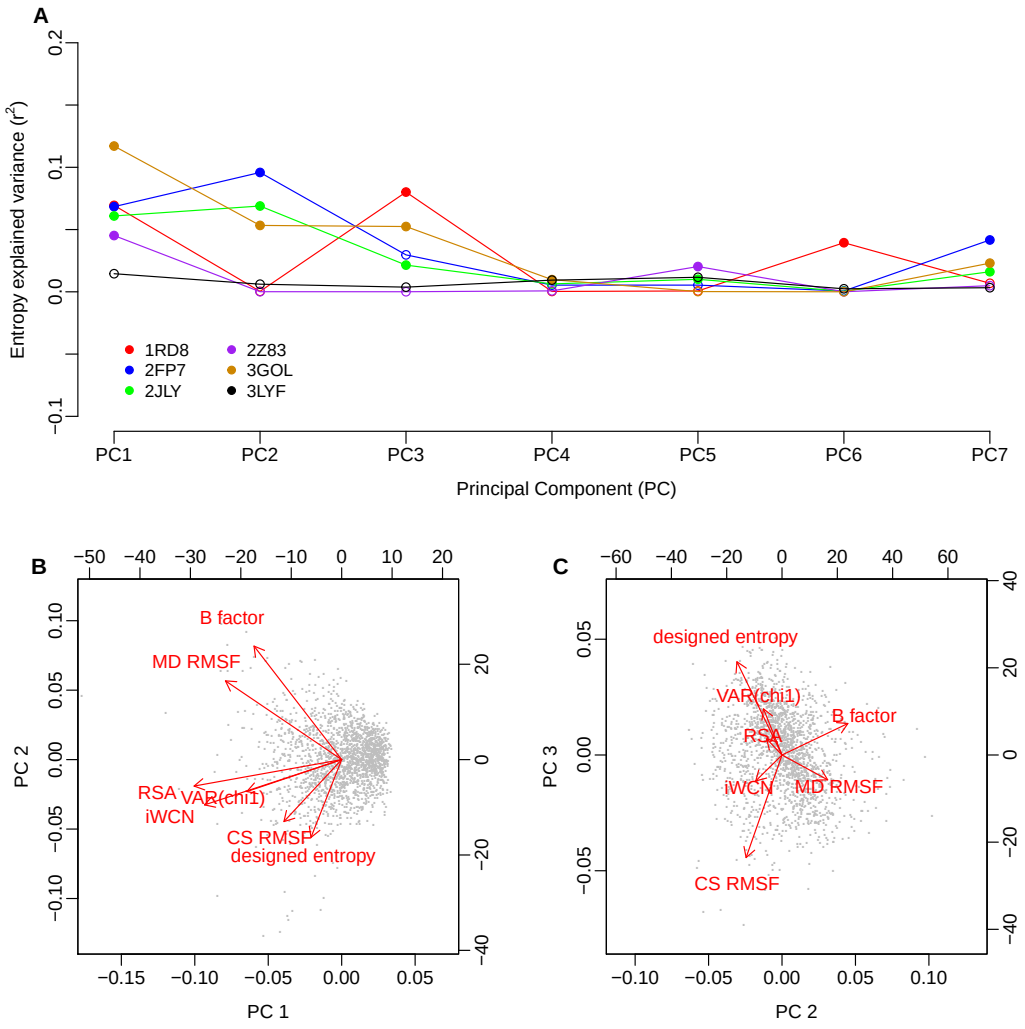

Figure 3.7: Principal Component (PC) Regression of sequence entropy against the structural variables, including CS RMSF. **(A)** Variance in entropy explained by each principal component. For most proteins, PC1 and either PC2 or PC3 show the strongest correlations with sequence entropy. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. **(B)** and **(C)** Composition of the three leading components. Red arrows represent the loadings of each of the structural variables on the principal components; black dots represent the amino acid sites in the PC coordinate system.

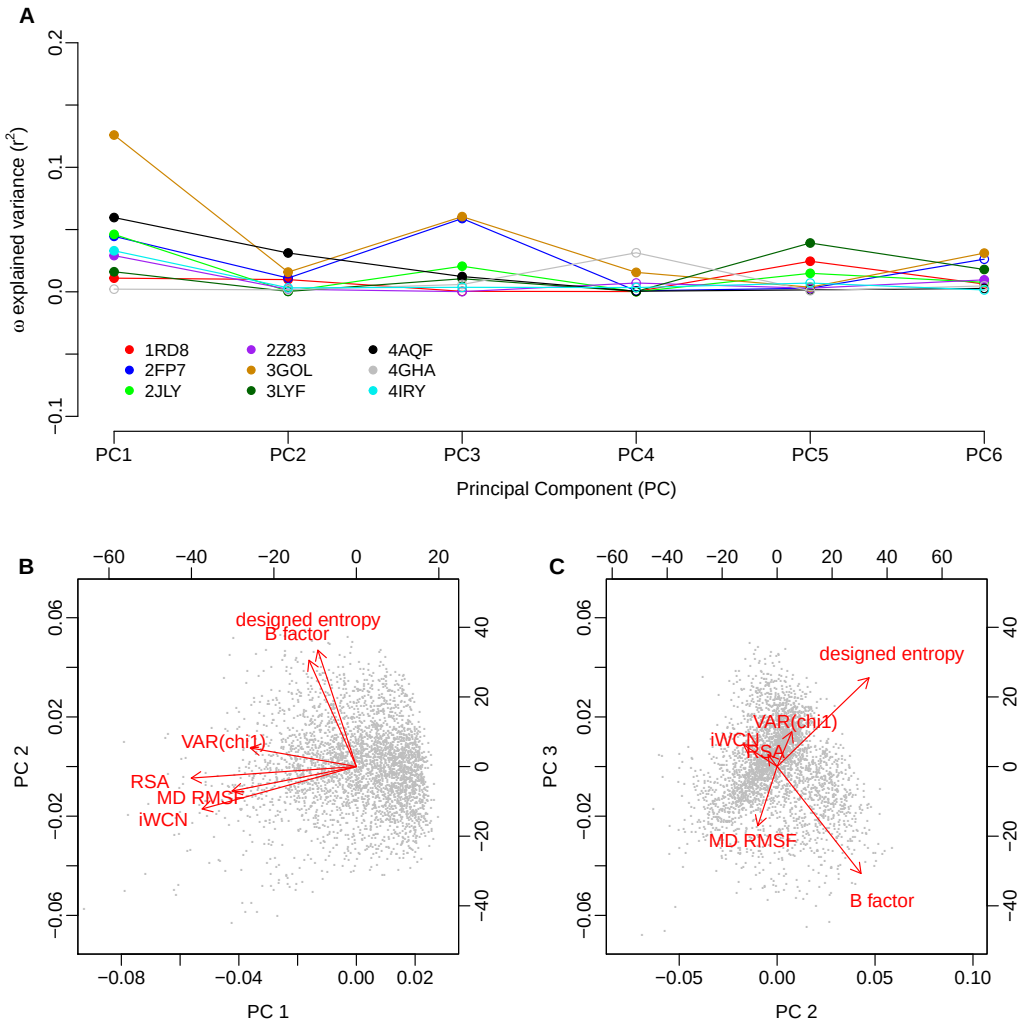

Figure 3.8: Principal Component (PC) Regression of ω against the structural variables. **(A)** Variance in ω explained by each principal component. For most proteins, PC1 and PC3 show the strongest correlations with ω. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. **(B)** and **(C)** Composition of the three leading components. Red arrows represent the loadings of each of the structural variables on the principal components; black dots represent the amino acid sites in the PC coordinate system. Note that parts B and C are identical to those shown in Figure 3.6.

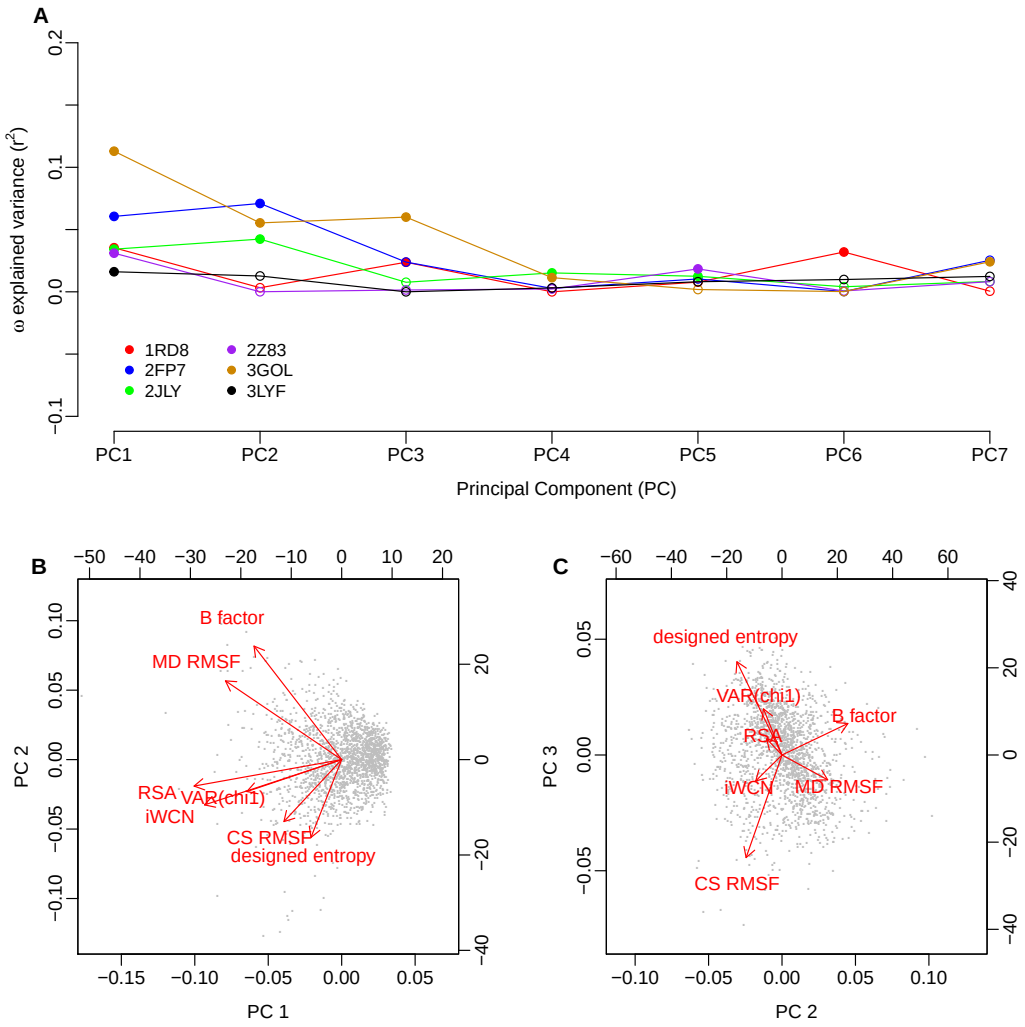

Figure 3.9: Principal Component (PC) Regression of ω against the structural variables, including CS RMSF. **(A)** Variance in ω explained by each principal component. For most proteins, PC1 and either PC2 or PC3 show the strongest correlations with ω. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. **(B)** and **(C)** Composition of the three leading components. Red arrows represent the loadings of each of the structural variables on the principal components; black dots represent the amino acid sites in the PC coordinate system. Note that parts B and C are identical to those shown in Figure 3.7.

in prior works. First, while viral sequences are abundant, their alignments may not be as diverged as alignments that can be obtained for sequences from cellular organisms. For example, our influenza sequences spanned only approximately one decade. Despite the high mutation rates observed in RNA viruses, the evolutionary variation that can accumulate over this time span is limited. This relatively lower evolutionary divergence makes resolving differences between more and less conserved sites much more difficult. Second, many viral proteins experience a substantial amount of selection pressure to evade host immune responses. The resulting positive selection on viral sequences may mask evolutionary constraints imposed by structure. For example, influenza hemagglutinin displays positive selection throughout the entire sequence, regardless of the extent of residue burial [10, 73, 74, 116]. However, the results I obtained here for viral proteins are broadly consistent with the results obtained earlier for cellular proteins [16, 23, 45, 131], indicating that viral proteins evolve under many of the same biophysical selection pressures that cellular proteins experience.

I have found here that correlations between sequence entropy and structural quantities were consistently higher than correlations between the evolutionary rate ratio ω and structural quantities. Surprisingly, in a recent study on cellular proteins, Yeh et al. (2014) [130] found that entropy performed worse than quantities assessing substitution rates. One possible explanation for this discrepancy is again our choice of viral sequences. Our sequence alignments almost certainly contained some polymorphisms, whereas the sequences

of Yeh et al. (2014) [130] likely did not. It is known that polymorphisms may diminish the reliability of ω estimates [56]. While the effect of polymorphisms on sequence entropy is not known, it seems plausible that entropy would be less sensitive to them than ω is. Alternatively, since viral proteins frequently experience positive selection, rate estimates may be confounded by this selection pressure and thus less reflective of constraints imposed by protein structure. By contrast, even under positive selection amino-acid distributions at sites would have to be consistent with the constraints imposed by the protein structure, and entropy would remain sensitive to these constraints.

I found that simple measures of buriedness or packing density, such as RSA or CN, were better predictors of evolutionary variation than was sequence variability predicted from computational protein design. In other words, simple quantities that can be obtained trivially from PDB structures performed better than a sophisticated protein-design strategy that makes use of an all-atom energy function and requires thousands of CPU-hours to complete. This result highlights that, even though computational protein design has yielded impressive results in specific cases [22,57,96], this approach remains limited in its ability to predict evolutionary variation. Similarly, I have previously found that flexible backbone design with Rosetta produced designs whose surface and core were too similar [45]. I attribute this discrepancy to either the solvation model or the model of backbone flexibility I used (Backrub; e.g., see [110]). The results I found here suggest that the model of backbone flexibility may indeed be the cause of at least some of the discrepancies between predicted

and observed site variability. In particular, in our PC regression analysis, the component in which designed entropy loaded opposite to B factor and MD RMSF generally had the second-highest predictive power for evolutionary variability, after the component representing buriedness/packing density. In sum, designed entropy was a better predictor for evolutionary sequence variability for sites with less structural flexibility compared to sites with more flexibility.

Even though RSA and CN remain the best currently known predictors of evolutionary variation, neither quantity has particularly high predictive power. One reason why predictive power may be low is that neither quantity accounts for correlated substitutions at interacting sites. Yet such correlated substitutions happen regularly. For example, covariation among sites encodes information about residue-residue contacts and 3D structure [9, 36, 46, 70], and evolutionary models that incorporate residue–residue interactions tend to perform better than models that do not [7, 94]. An improved predictor of evolutionary variation would have to correctly predict this covariation from structure. In principle, computational protein design, which takes into consideration the atom-level details of the protein structure, should properly reproduce covariation among sites. However, a recent analysis showed that there are significant limitations to the covariation that is predicted [82]. In addition, covariation in designed proteins is quite sensitive to the type of backbone variation modeled during design, and improved models of backbone flexibility may be required for improved prediction of covariation among sites [82].

Chapter 4

Structural Determinants of Sequence Evolution in Enzymatic proteins

4.1 Introduction

A variety of site-specific structural characteristics have been proposed over the past decade to predict protein sequence evolution from structural properties. Among the most important and widely discussed are the Relative Solvent Accessibility (RSA) [11,12,25,33,75–77,90,102,104,108,132,133], Contact Number [6,39,43,62,69,77,93,104,132,133], measures of thermodynamic stability changes due to mutations at individual sites in proteins [19,123], and measures of local flexibility, such as the Debye-Waller factor (hereafter B factor) [62,104,107] or flexibility measures based elastic network models [66] and Molecular Dynamics (MD) simulations [104].

Although structural characteristics have been individually extensively studied and explored with regards to their association with sequence evolution, it is yet unknown whether these seemingly independent quantities are merely different manifestations of a more fundamental underlying characteristics of individual sites in proteins or each influence the sequence evolution independently. It is perceivable that quantities such as B factor, RSA, and

CN, all serve as a proxy measures of local packing density of individual sites in proteins, or the local flexibility of individual amino acids. Franzosa & Xia (2009) [25] used a variety of structural variables representing the local packing density to show that RSA is the key determinant of sequence evolution with packing density having only peripheral influence. Recently however, Huang et al. (2014) [43] have argued, through an extensive mathematical formulation within the framework of Elastic Network Models, for the local packing density as the dominant factor in sequence variability patterns in contrast to RSA and local flexibility measures.

It is notable that the site-specific flexibility is often represented by C_α atomic B factor, a quantity that is not necessarily an unbiased measure of the amino acid flexibility as a whole in a given site in protein. A more accurate measure of amino acid flexibility requires the calculation of accessible free volume to each site in protein structure. An estimate of the accessible volume for each site in protein can be generally obtained through a quantity widely known as Contact Number introduced and discussed by several authors [62]. In its simplest mathematical form, the Contact Number for a given site in protein is defined as the number of amino acids within a fixed radius r of neighborhood around it [25]. Individual sites are generally represented by the coordinates of C_α backbone atoms for the calculation of CN. A major problem with the traditional definition of contact number however, is the existence of the arbitrary parameter r in the definition of CN. There is no consensus on the optimal value of this cutoff distance, although it is typically chosen in the

range $7\mathring{A}$ to $13\mathring{A}$ [25, 64].

In an attempt to provide a more general definition of CN, some studies [64] have already suggested an alternative definition known as the Weighted Contact Number (WCN): For a given site i in a protein of length N, WCN_i is defined as the sum of the inverse-squared of distances between the amino acid of interest and all other sites in protein,

$$WCN_i = \sum_{j \neq i}^{N} {r_{ij}}^{\alpha=-2}, \tag{4.1}$$

Although WCN is in general a better predictor of C_α atomic B factor and site-specific sequence variability, the proposed definition of WCN still involves an adjustable free parameter, the exponent of the power-law kernel, which is typically fixed to $\alpha = -2$ as shown in Eqn 4.1 [126]. Moreover, no physical model has been so far proposed to support the power-law kernel used in the definition of WCN and the specific value of exponent often used.

Motivated by the existing gaps in the current understanding of the role of flexibility and other structural properties on sequence-structure relations in proteins, here I propose and derive a new set of site-specific structural properties which, unlike CN and WCN, their definitions does not involve any free parameters, while performing equally well or better than all previously-considered structural quantities in predicting protein sequence evolution. This is done by employing tessellation methods from the field of computational geometry and Condensed Matter Physics to calculate several new characteristics

of sites in proteins, which can serve as proxy measures of local packing density and site-specific flexibility. Contrary to what is currently perceived about the role of flexibility in sequence variability [43], I show that the newly calculated flexibility measures outperform many of previously studied structural properties, such as RSA and the traditional definitions of Contact Number and the Weighted Contact Number (WCN), in predicting sequence evolution at residue level.

Furthermore, for structural properties that are calculated based on a set of representative site coordinates, I show that the choice of the geometric average of the side chain atomic coordinates instead of the traditional choice of C_α atomic coordinates, always results in significantly better predictions of site-specific sequence evolution. Similar improvements in correlations with different ste-specific structural properties and sequence variability measures are also observed if the average of side chain B factors, instead of C_α atomic B factor, is used as a proxy measure of site flexibility.

I also show that the original kernel proposed for the definition of Weighted Contact Number by [64] and supported further by [126] and extensively used in other studies, has no significant advantage whatsoever in predicting site-specific flexibility measures (e.g., B factor) or the sequence variability, when compared to other possible types of kernels. A discussion of the methodology used in this work, the results and implications of our findings on the energy landscape of proteins and sequence-structure relations will be presented in the following sections.

All data including a list of 209 proteins and their properties together with Python, R and Fortran codes written for data reduction and analysis are publicly available to view and download at `https://github.com/shahmoradi/cordiv`.

4.2 Protein Dataset and Structure/Sequence Variability Measures

The entire analyses and results presented in this work are based on a dataset of 209 monomeric enzymes [19,133] randomly picked from the Catalytic Site Atlas 2.2.11 [87] with protein sizes in the sample ranging from 95 to 1287 amino acids, including representatives from all six main EC functional classes [121] and domains of all main SCOP structural classes [79]. To assess the evolutionary rates at the amino acid level for each protein, a set of up to 300 homologous sequences were used [133] for each protein from the *Clean Uniprot* database following the ConSurf protocol [2, 30]. Sequence alignments were then constructed using amino-acid sequences with MAFFT [53], specifying the auto flag to select the optimal algorithm for the given data set, and then back-translated to a codon alignment using the original nucleotide sequence data.

The alignments were then used to calculate the site-specific sequence variability for each individual protein in dataset. Two independent methods were used for the assessment of sequence variability. First, the respective sequence alignment for each structure in the dataset and phylogenetic tree were

used to infer the site-specific evolutionary rates ($\omega = \mathrm{d}N/\mathrm{d}S$) with Rate4Site, using the empirical Bayesian method and the amino-acid Jukes-Cantor mutational model [72], hereafter abbreviated as *r4sJC*. The quantity ω is the ratio of the number of non-synonymous substitutions per non-synonymous site ($\mathrm{d}N$) to the number of synonymous substitutions per synonymous site ($\mathrm{d}S$), which can be used as an indicator of selective pressure acting on a protein-coding gene. A synonymous substitution refers to the evolutionary substitution of one nucleotide base with another in the codon sequence of the protein, such that the resulting amino acid in the specific site of interest in protein is not altered, whereas a non-synonymous nucleotide substitution in the codon sequence of protein alters the amino acid sequence of the protein in the specific site of interest.

In addition to site-specific evolutionary rates, the Shannon entropy (H_i) – the sequence entropy [105] – was also calculated at each alignment column i according to Eqn. 2.1, based on the assumption that the occurrence of each of the 20 amino acids is equally likely at any given site in the alignments.

The Solvent Accessible Surface Area (SASA) were calculated using DSSP software [48] for individual amino acids in all sites in proteins using a spherical probe of radius $\sim 1.5\mathring{A}$ representing water molecule. Since the 20 naturally occurring amino acid molecules come in different sizes, it is also necessary to normalize the SASA values of individual amino acids to the their corresponding *maximum solvent accessibility*. Here the SASA values from DSSP were normalized to the computationally calculated maximum SASA values

of [118] to obtain the Relative Solvent Accessibility (RSA) for all individual sites in all proteins.

A measure of thermodynamic stability changes due to amino acid substitutions at individual sites in proteins can be defined and obtained following the stability threshold model of Bloom et al. (2006) [6], which was also recently further studied by Echave et al. (2014) [19]. Based upon this model which was extensively described in Chapter 2, a quantity $\Delta\Delta G$ *rate* (or it ddG rate) was derived for each individual site in all proteins in dataset. A high ddG rate for the i^{th} site in a protein indicates a high stability of the site and the overall conformation of the protein to perturbations caused by substitution of the amino acid residing the site.

As a measure of local flexibility or fluctuation in different parts of the protein structure the temperature, factors (*B factor*) for all atoms in PDB files were extracted (c.f., Chapter 2). Although, B factor is an atomic measure of flexibility and fluctuation in proteins, the backbone C_α B factor has become a very popular proxy measure of amino acid flexibility in the studies of protein dynamics and benchmarking of different Elastic Network Models of proteins. Alternatively, the site-specific fluctuation could be calculated from MD simulations. This was however impossible for this study for the large dataset of 209 proteins considered in this work.

4.3 Voronoi Partitioning of Protein's Structure

There is already extensive body of literature on the applications of different methods of structural partitioning in the studies of protein structure and its prediction from sequence [29, 92]. The Voronoi tessellation and its dual graph, the Delaunay triangulation, have particularly attracted much attention in the studies of protein internal structure and development of empirical potentials [125, 134, 135]. For a given a set of centroid points (seeds) in 3-dimensional Euclidean space, the simplest and most familiar case of Voronoi tessellation divides the space into regions, called *cells*, such that the cell for each centroid point consists of every region in space whose distance is less than or equal to its distance to any other centroid points (Figure 4.1).

In the context of protein studies, the atomic coordinates of C_α backbone atoms have been widely used as the set of Voronoi seeds to partition the 3D structure of protein according to Voronoi tessellation. An example of Voronoi tessellation of protein structure in two dimensions (PDB ID: *1LBA*) is shown in Figure 4.1. The properties of individual cells resulting from tessellation are then used to obtain a wide range of information on protein structure, energy landscape or protein–protein interactions.

Here in this work, the simplest and most widely used definition of Voronoi tessellation described above is applied on a dataset of 209 monomeric enzymes. We use VORO++ software [98] to calculate the relevant Voronoi cell properties of all sites in all proteins in the dataset. Among the most important properties are the length of the cell edges, cell area and volume, number of faces

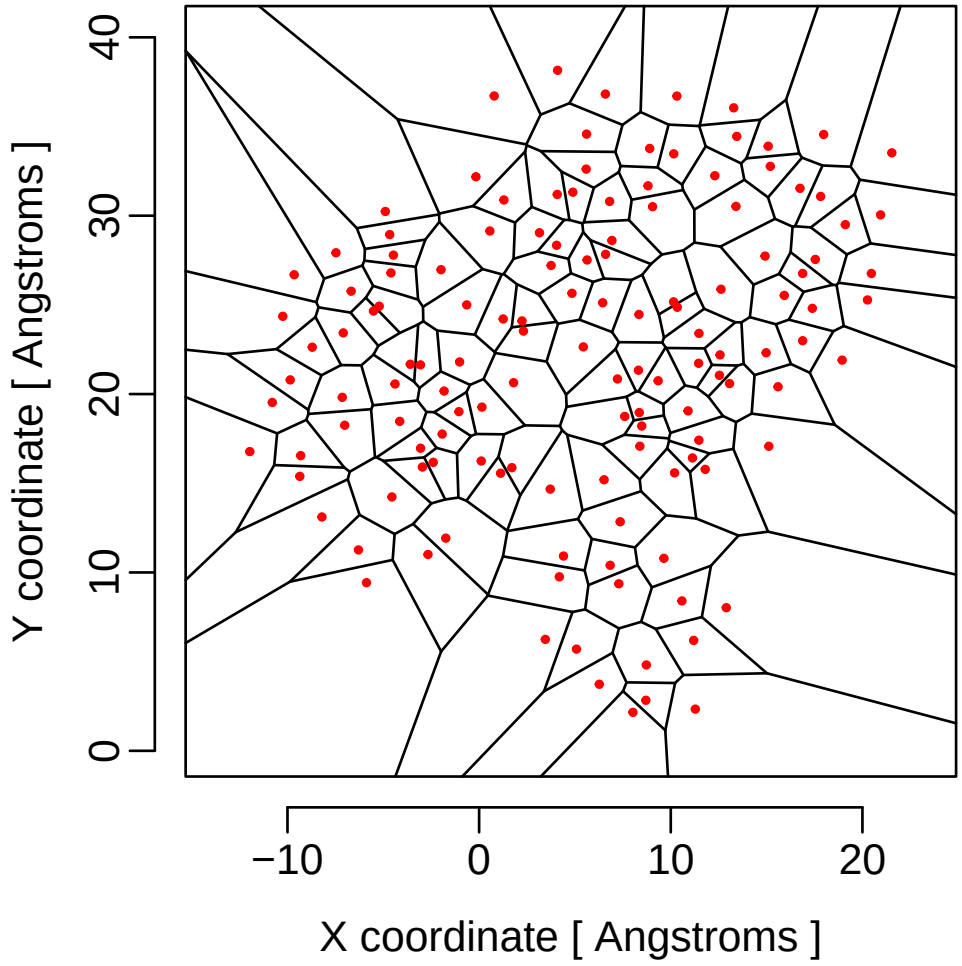

Figure 4.1: An Example 2-dimensional Voronoi diagram for bacteriophage T7 lysozyme (Protein Data Bank ID '1LBA'). The red dots represent the backbone C_α atoms projected on the X–Y plane, used as cell seeds in Voronoi tessellation.

of each cell, the cell eccentricity defined as the distance between the cell's seed and the geometrical center of the cell. A measure of the cell *eccentricity* can be also obtained by finding the distance between the cell seed and geometrical center of the cell. In addition, the cell *sphericity* can be calculated as a measure of the cell's *compactness* defined as,

$$\Psi = \frac{\pi^{\frac{1}{3}}(6V)^{\frac{2}{3}}}{A}. \tag{4.2}$$

in which V & A stand for the volume & area of the cell respectively. For a perfectly spherical cell, $\Psi = 1$, while it becomes zero for a 2-dimensional object that has no volume but only surface area.

4.3.1 Voronoi Cell Area and Volume as Proxy Measures of Local Packing Density and Flexibility in Proteins

In order to assess the prediction power of site-specific variables derived from Voronoi tessellation, first the geometric centers of all side-chains for each of the proteins in dataset were calculated and used as the seeds of Voronoi polyhedra. Figure 4.2 depicts the distributions of the Spearman's correlation coefficients of five most important Voronoi cell characteristics with site-specific evolutionary rates (ER). It is notable that all cell characteristics in the plot correlate positively with ER, except the cell sphericity which is always negatively correlated with ER and other Voronoi cell properties. In general, it is observed that the cell surface area has the best prediction power compared to other cell characteristics, followed by the cell volume, cell eccentricity as

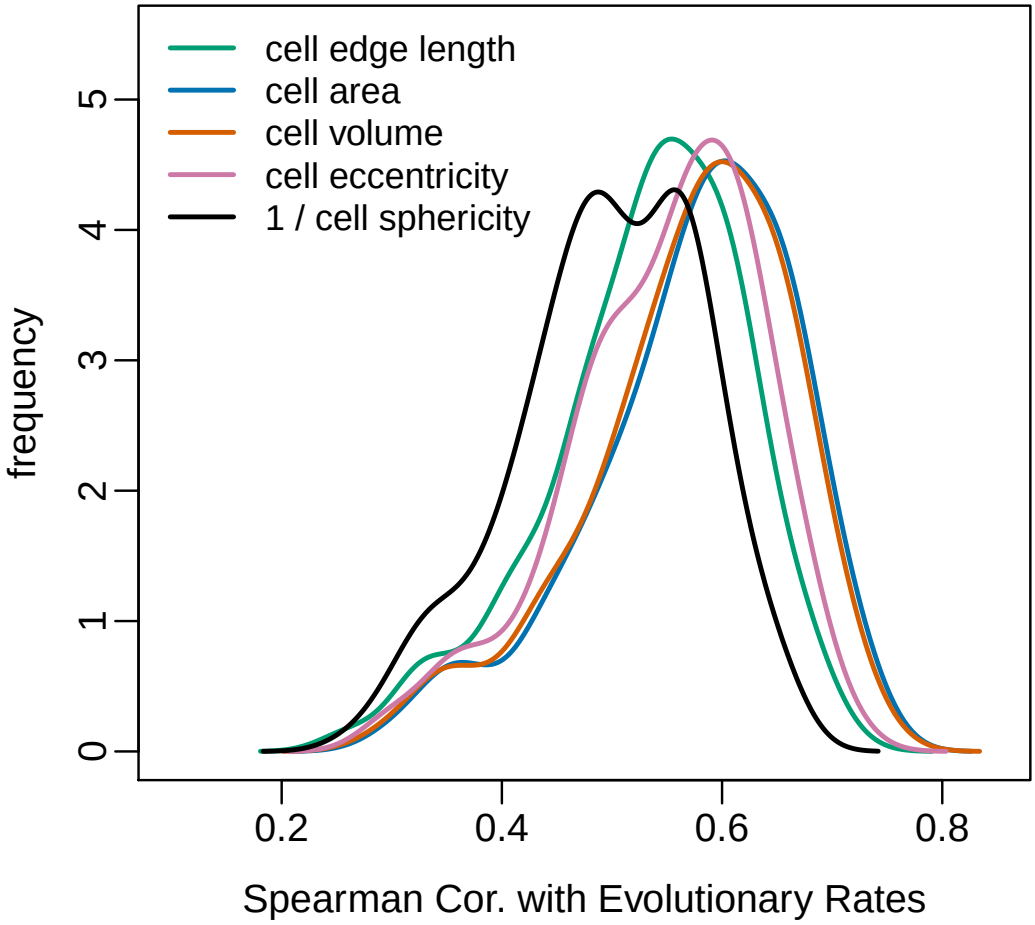

Figure 4.2: A comparison of the prediction power of different Voronoi cell characteristics about site-specific evolutionary rates (ER). Note that all cell characteristic correlate positively with ER, except sphericity which strongly negatively correlates with ER.

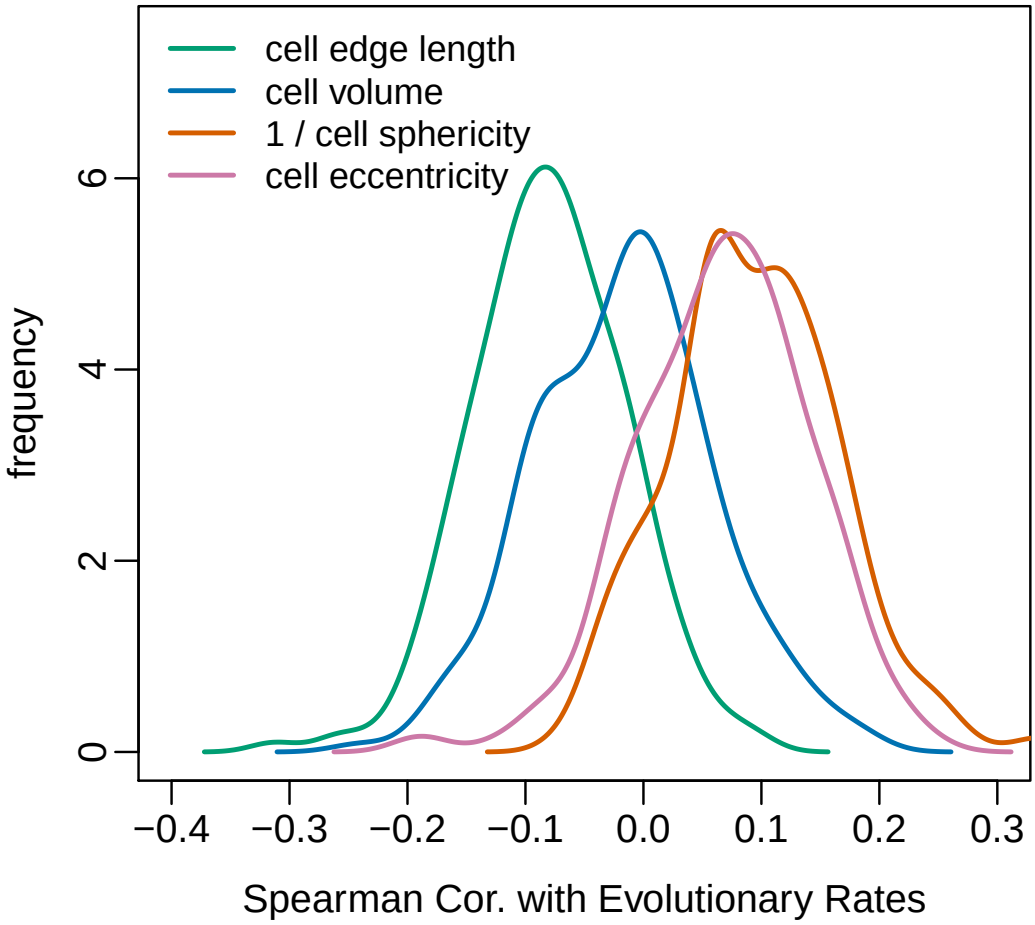

Figure 4.3: The partial correlation strengths of the same Voronoi cell characteristics with sequence evolutionary rates while controlling for the cell area.

defined in previous section, cell's total edge length, and the cell sphericitiy. The cell properties are also strongly correlated with each other. Although the Voronoi cell volume is the second best correlating variable with ER, it exhibits no significant independent correlation with ER once we control for the cell area, with the median of its distribution centered at ~ 0.0, as illustrated in Figure 4.3. Conversely, the cell sphericity and eccentricity both exhibit median partial correlations of ~ -0.1 & ~ 0.07 with ER respectively, when the contribution from the Voronoi cell area is controlled. In conclusion, the cell area, volume, and edge length appear to almost represent the same property of the Voronoi cell. Other Voronoi cell characteristics, such as the number of vertices, faces and edges of the cell also tend to correlate weakly with sequence evolutionary rates. These cell characteristics are however, discrete (integer) quantities and in general have a limited range.

Not shown here for brevity, almost identical results to the above are obtained if sequence entropy as defined by Eqn. 2.1 were used in place of sequence evolutionary rates. The use of sequence entropy however, generally results in weaker correlation strengths due to the discreteness and limited range inherent in the definition of sequence entropy.

One potential caveat with Voronoi tessellation of finite structures in Euclidean space is the *edge effects*. Sites that are close to the surface of protein are often associated with Voronoi cells that are bounded by the cubic box containing the protein (Figure 4.1). Here to ensure that these edge effects do not influence the observed sequence-structure correlations, the open cells –

i.e., cells that are partially bounded and closed by the cubic box containing the protein – are identified in all proteins by examining the variations in individual cell volumes upon changing the size of the cubic box containing the protein to a given extreme value. The open cells in individual proteins are then ranked by the fraction of volume changes observed upon changing the box size and then normalized to the the largest volume observed among closed cells. It should be noted that the specific extreme value chosen for the box sizes of the proteins or the rank ordering of the open cells does not have any influence on the resulting correlation strengths, since the Spearman's ρ by its definition is a rank correlation coefficient.

4.4 Average Side Chain coordinates as the Best Representation of Protein 3D Structure

Depending on the choice of the Cartesian coordinates used, there exist degeneracy in the definition of some site-specific structural variables. For example, the quantity WCN is generally calculated from the coordinates of C_α atoms in the 3-dimensional structure of protein. The choice of C_α coordinates is however mainly driven by convenience in WCN calculation and there is no reason to believe this set of atomic coordinates is the best representative of individual sites in proteins. Indeed, some earlier works have already suggested the use of center-of-mass of side chain coordinates to represent the 3D structure of protein [111]. More recently, Marcos & Echave (2014) [69] have also shown that WCN calculated from side-chain center-of-mass coordinates gener-

ally result in significantly better correlations of WCN with sequence variability measures.

Despite the highly popular choice of C_α atomic B factor as a proxy measure of residue flexibility [38], same definition degeneracy also exists on choice of atomic B factors that are used to represent site-specific flexibility. In addition to WCN and B factor, there is also ambiguity as to which set of residue atomic coordinates best represent individual sites in proteins for the generation of Voronoi polyhedra.

Here in this work, all possible choices of the representative set of atomic coordinates are considered in order to identify which set of atomic coordinates best represents individual sites for the calculation of WCN, B factor, and Voronoi cells. Depending on the set of atomic coordinates that represent the protein structure, there are at least 7 different measures of each individual site-specific structural properties, such as the Weighted Contact Number, B factor and Voronoi cell properties. These include the set of coordinates of all backbone atoms (N, C, C_α, O) and the first heavy atom in the amino acid side chains (C_β). In addition, representative coordinates for each site in protein are calculated by averaging over the coordinates of all heavy atoms in the side chains. Also calculated is a representative coordinate for each site by averaging over all heavy atom coordinates in the side chain and the backbone of the amino acid together. In rare cases where the side chain C_β atom had not been resolved in the PDB file or the amino acid lacked C_β (e.g., Glycine), the C_β coordinate for the specific amino acid were replaced with

the coordinate of the corresponding C_α atom in the same amino acid. The resulting Spearman's correlation strengths of site-specific evolutionary rates, sequence entropy, $\Delta\Delta G$ rate, Relative Solvent Accessibility (RSA), amino acid hydrophobicity, and Hydrogen bond energy with different measures of WCN, B factor, and Voronoi cell area are depicted in the plots of Figures 4.4, 4.6, and 4.5 respectively, for different sets of atomic coordinates used in the calculations. The hydrophobicity scales of amino acids residing in individual sites in proteins were taken from [40]. Other hydropobicity scales were also considered [58,124], however similar results are obtained for all.

For the measure of local packing density in proteins (the Weighted Contact Number) we find that among all possible set of coordinates, the average over coordinates of all heavy atoms of each individual side chain results in WCN values that show the strongest correlation strength with other structural and sequence properties, such as RSA, Voronoi cell properties, sequence entropy, and evolutionary rates. Specifically, WCN from average side chain coordinates outperforms WCN based on C_α coordinates in predicting RSA, $\Delta\Delta G$ rate, sequence entropy and evolutionary rates with median Spearman correlation differences of 0.09, 0.10, 0.07 & 0.08, respectively (Figure 4.4).

For the measure of local flexibility in proteins (B factor) we similarly find that among all 7 representative measures of site B factors, the average of B factor values over all heavy atoms of each individual side chain results in the best correlations with other structural and sequence properties. Specifically, the average side chain B factor outperforms the commonly used C_α B factor

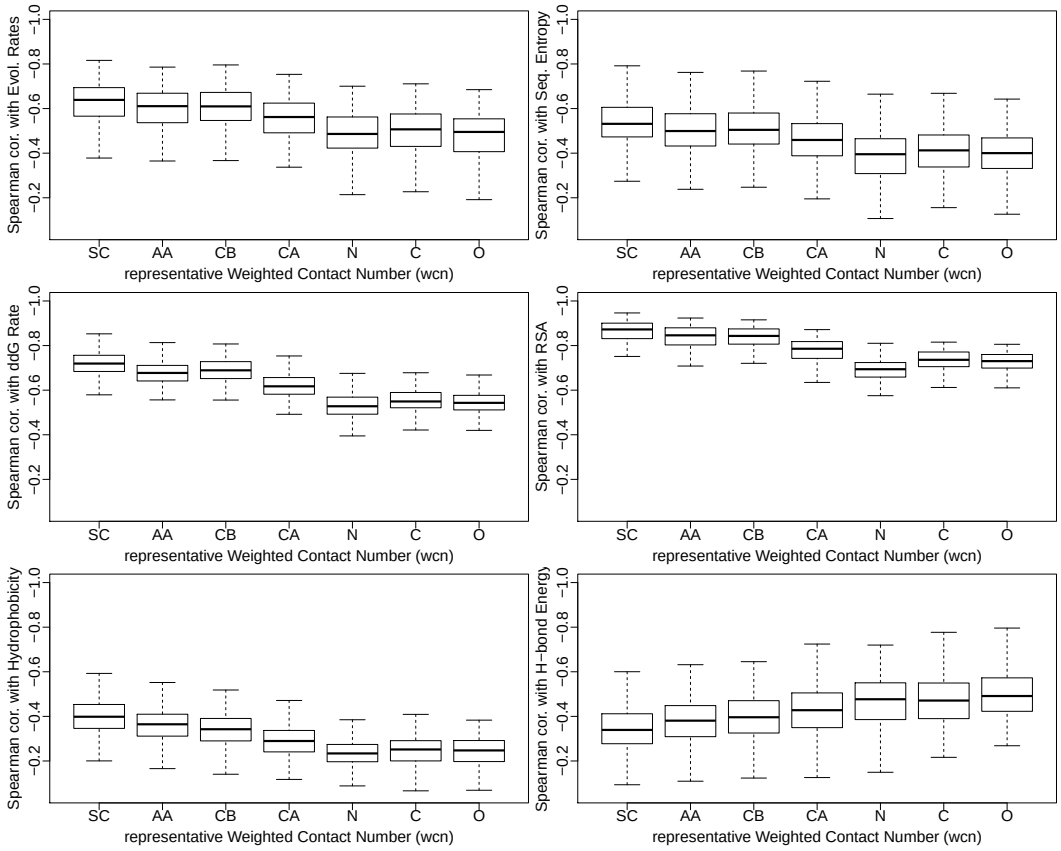

Figure 4.4: A comparison of the correlation strength of 6 different measures of Weighted Contact Number (WCN) with 6 coordinate-independent structural or sequence properties for 209 proteins in dataset. The contact numbers, WCN, are calculated using 6 sets of atomic coordinates: *SC, AA, CB, CA, N, C, O*, used as different representations of individual sites in proteins. The two labels *SC* & *AA* stand respectively for the geometric average coordinates of the Side Chain (SC) atoms and the entire Amino Acid (AA) atoms, excluding hydrogens.

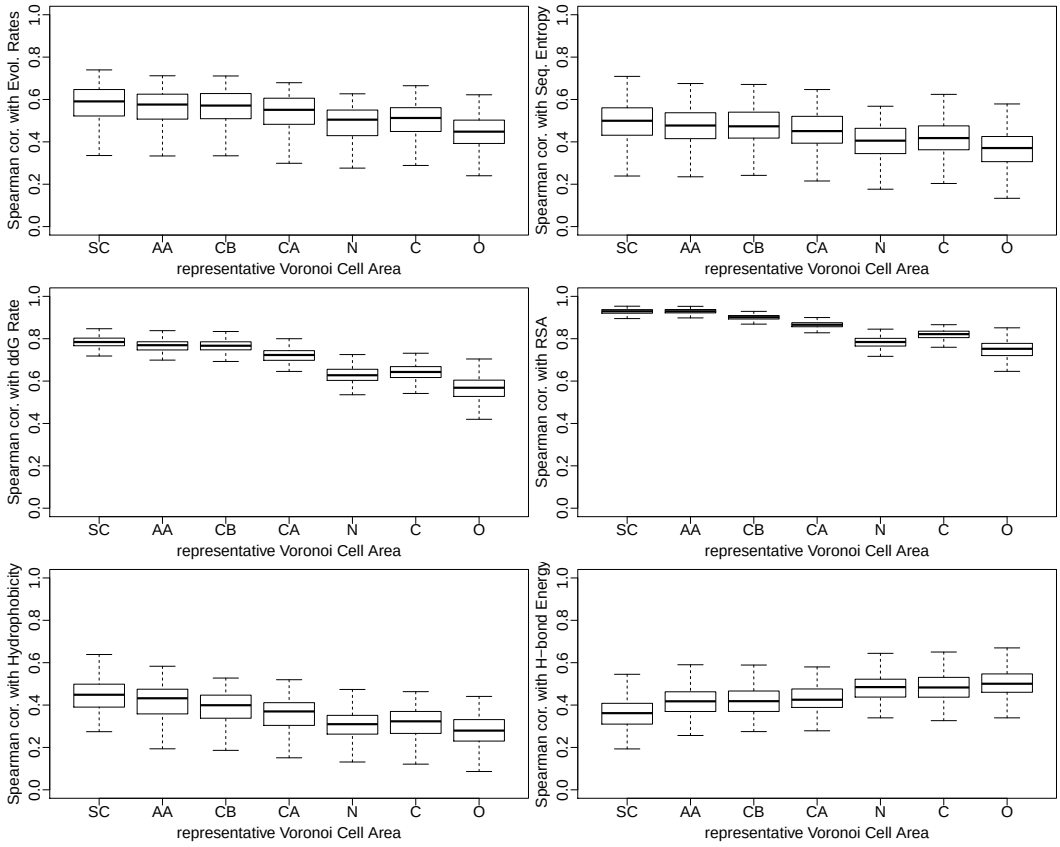

Figure 4.5: A comparison of the correlation strength of 6 different measures of Voronoi cell areas with 6 coordinate-independent structural or sequence properties for 209 proteins in dataset. The Voronoi cells are generated using 6 sets of atomic coordinates: *SC, AA, CB, CA, N, C, O*, used as different representations of individual sites in proteins. The two labels *SC* & *AA* stand respectively for the geometric average coordinates of the Side Chain (SC) atoms and the entire Amino Acid (AA) atoms, excluding hydrogens.

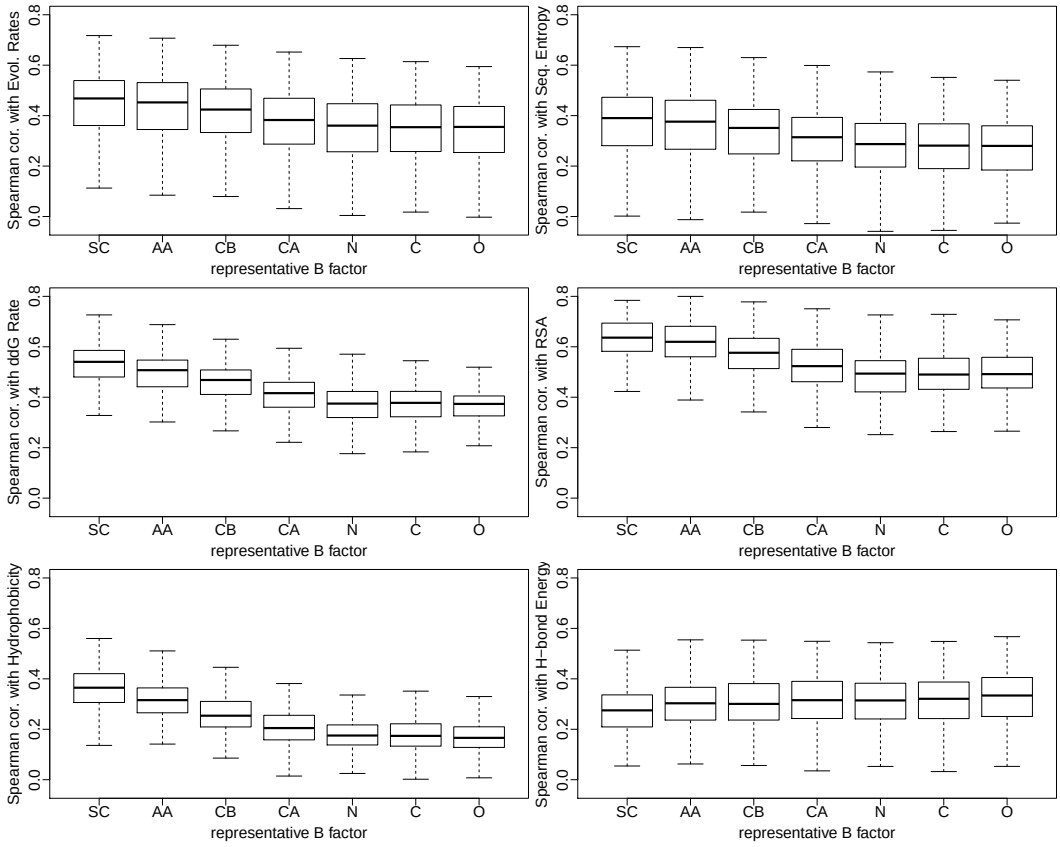

Figure 4.6: A comparison of the correlation strength of 6 different measures of B factor with 6 coordinate-independent structural or sequence properties for 209 proteins in dataset. Shown on the horizontal axes, are the 6 representative atomic B factors: *SC, AA, CB, CA, N, C, O* used as flexibility measures of individual sites in proteins. The two variables SC & AA stand respectively for the average B factor of all Side Chain (SC) atoms and the entire Amino Acid (AA) atoms, excluding hydrogens.

in predicting RSA, $\Delta\Delta G$ rate, sequence entropy and evolutionary rates by a median Spearman correlation difference of 0.11, 0.12, 0.08 & 0.09, respectively (Figure 4.6).

Similar to WCN and B factor, the Voronoi cell properties, most importantly the cell surface area, volume, edge length, eccentricity and the cell sphericity also correlate best with other structure and sequence properties, only if the geometric average of side chain coordinates are used as the seeds of Voronoi cells. Specifically, cell area from average side chain coordinates outperforms cell area based on C_α coordinates in predicting RSA, $\Delta\Delta G$ rate, sequence entropy and evolutionary rates with median Spearman correlation differences of 0.04, 0.06, 0.04 & 0.04, respectively (Figure 4.5).

It is notable that the standard deviations of the difference distributions for all three quantities: WCN, B factor, and Voronoi cell area, are an order of magnitude smaller than the observed differences, implying that the correlation coefficients for all proteins in dataset uniformly translate to higher values by moving from C_α atomic coordinates to the geometric centers of the side chains, regardless of the strength of the correlation coefficients.

4.5 Discussion

Throughout this work, a comprehensive analysis and comparison of the main structural determinants of sequence variability was carried out, using a dataset of 209 monomeric enzymes. Examples of sequence–structure relations include the correlations of measures of evolutionary rates such as *r4sJC* used

in this work and sequence entropy, with measures of residue Contact Number, Relative Solvent Accessibility (RSA), and $\Delta\Delta G$ rate as defined in Chapter 2 (see also Echave et al. (2014) [19]), which is essentially a proxy measure of the stability of protein's native conformation upon substitution of amino acids in individual sites in proteins. In addition, we have derived new site-specific characteristics from the Voronoi Tessellation of protein 3D structures, that are capable of explaining sequence variability equally well or better than several previously considered structural quantities, such as B factor, RSA, $\Delta\Delta G$ rate, and the traditional definitions of contact number and the weighted contact number (WCN) using C_α atomic coordinates (e.g., Figures 4.7 & 4.8).

One potential caveat with Voronoi tessellation of finite structures in Euclidean space is the *edge effects*. However, based on the results of the analysis presented in Section 4.3, the *edge effects* due to Voronoi tessellation appear to have $\lesssim 0.01$ influence on the observed sequence-structure correlations in the dataset of 209 proteins considered in this work. Similar conclusions are reached if the open cells were alternatively ranked by different criteria such as the fractional changes in cell area (vs. cell volume) upon changing the box size. The Voronoi cell characteristics, in particular cell volume and cell area can be safely used in predicting sequence variability without recourse to corrections for edge effects. An exception however is cell sphericity as defined in Eqn. 4.2, which turns out to behave differently for open and closed cells. This is well illustrated in the adjacent averaging plots of Figure 4.9 in which the behavior open and closed Cell characteristics, averaged over all sites in

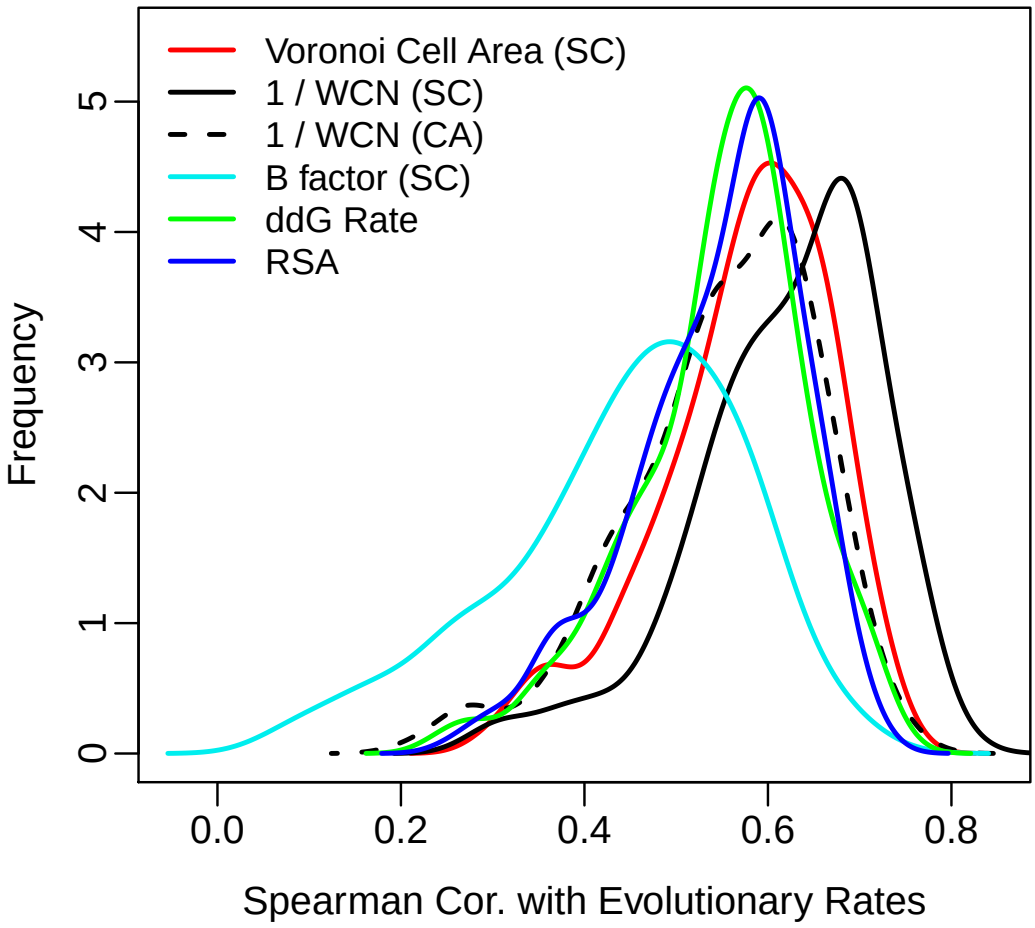

Figure 4.7: A comparison of the prediction power of five structural variables about site-specific evolutionary rates (ER). All structural quantities correlate positively with ER, with the exception of Weighted Contact Number (WCN) which correlates negatively. For better illustration however, the Spearman's correlation coefficient (ρ) of the inverse of WCN with ER are shown in the Figure. Note that the Spearman's ρ is a rank correlation coefficient, meaning that the use of inverse WCN only changes the sign and not the magnitude of ρ. The abbreviation *SC* refers to the use of average Side-Chain coordinates or average Side-Chain B factor wherever used, and *CA* refers to the use of backbone C_α atomic coordinates for representation of individual sites in proteins. The paired t-test for the significance of the the difference in the observed distributions of correlation strengths are available online in the repository of the project.

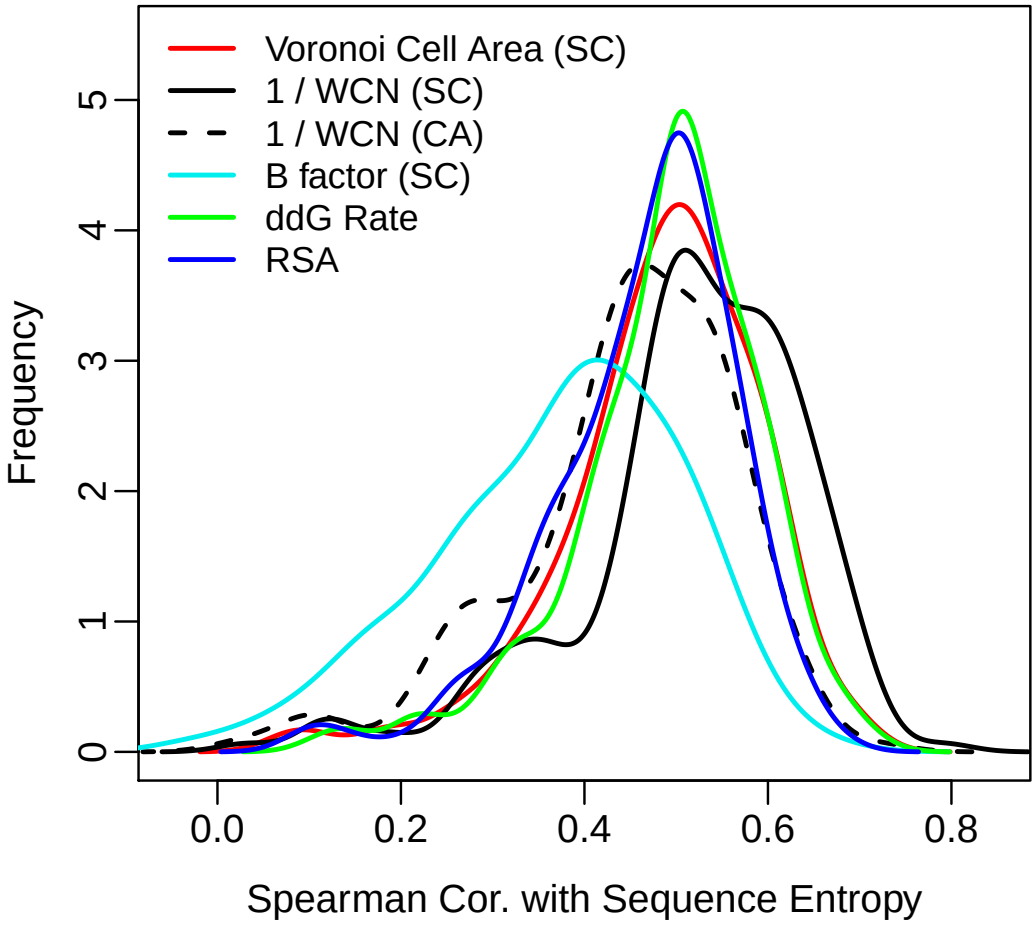

Figure 4.8: A comparison of the prediction power of five structural variables about site-specific Sequence Entropy (SE). All structural quantities correlate positively with SE, with the exception of Weighted Contact Number (WCN) which correlates negatively. For better illustration however, the Spearman's correlation coefficient (ρ) of the inverse of WCN with ER are shown in the Figure. Note that the Spearman's ρ is a rank correlation coefficient, meaning that the use of inverse WCN only changes the sign and not the magnitude of ρ. The abbreviation *SC* refers to the use of average Side-Chain coordinates or average Side-Chain B factor wherever used, and *CA* refers to the use of backbone C_α atomic coordinates for representation of individual sites in proteins.

all proteins in our dataset, are plotted against the *normalized* sequence evolutionary rates. For comparison, Figure 4.10 depicts the general behavior of the normalized site-specific evolutionary rates versus site-specific sequence entropy, $\Delta\Delta G$ rate, RSA, WCN, average Side-Chain B factor, Hydrogen bond strengths.

Also calculated in this work, were the site-specific structural quantities using different sets of atomic coordinates representing individual sites in proteins. These include the weighted contact number, the Voronoi cell characteristics, and representative site-specific B factor. All observations clearly demonstrate that individual sites in proteins are best represented by the average properties of the side chains of amino acids in the corresponding sites. In particular, the strength of structure-structure and sequence-structure correlations decrease monotonically by moving from side chain to backbone atoms.

An exception to this general pattern is the correlation of the hydrogen-bond energies of sites with other site-specific structural properties. In general, average site-specific H-bond energies correlate more strongly with representative B factors, contact number, and Voronoi cell characteristics, if calculated using the backbone Oxygen atom coordinates in individual sites, instead of average side chain coordinates. This is well illustrated in the bottom-right plots of Figures 4.5, 4.4, & 4.6. The observed monotonic increase in the correlation strengths with H-bond energies from side chain to backbone O atom can be explained away knowing that the backbone Oxygen atom is responsible for virtually all Hydrogen bonds in proteins. The influence of individual atoms on

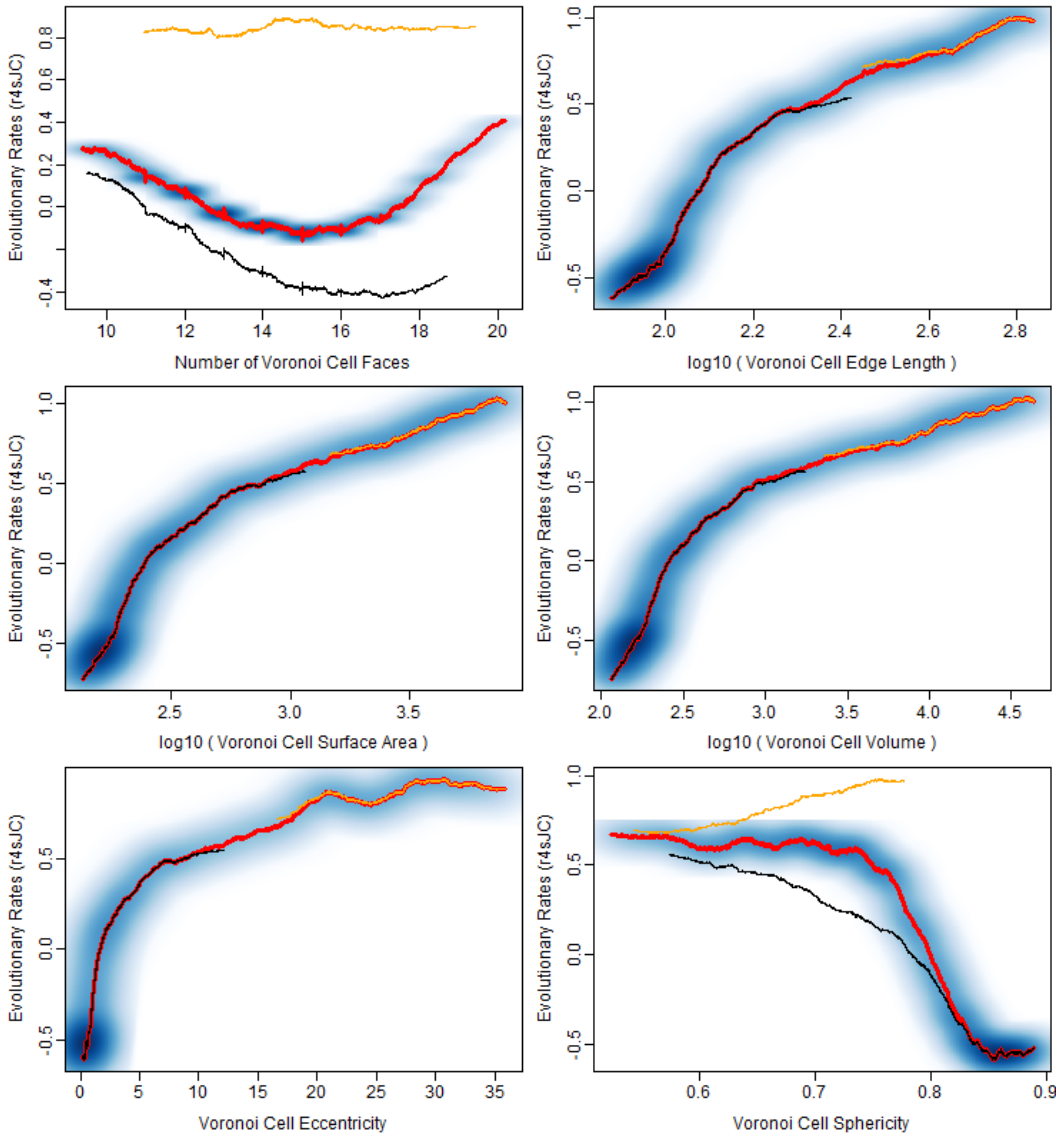

Figure 4.9: General behavior of Voronoi cell characteristics versus normalized site-specific evolutionary rates among all sites in all 209 proteins in dataset. The red curves in each plot is obtained by adjacent-averaging of every 3000 sites. The black & orange curves represent respectively the general behaviors of closed & open Voronoi cell characteristics. The blue-shaded area in each plot is a heat map indicating the overall concentration of 75755 sites in all 209 proteins along the horizontal axis.

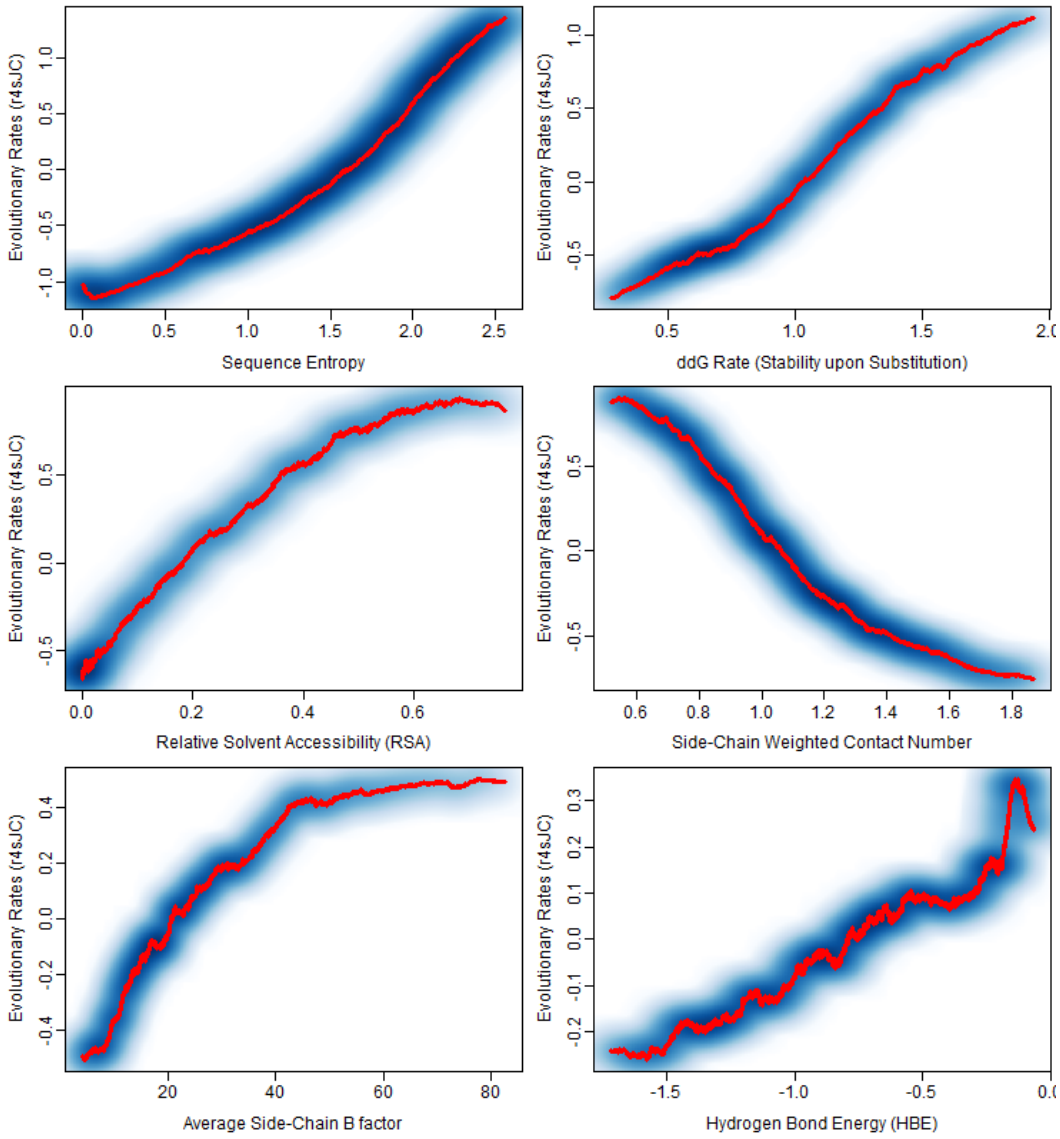

Figure 4.10: General behavior of site-specific structural characteristics versus site-specific evolutionary rates among all sites in all 209 proteins in dataset. The red curves in each plot is obtained by adjacent-averaging of every 3000 sites. The blue-shaded area in each plot is a heat map indicating the overall concentration of 75755 sites in all 209 proteins along the horizontal axis.

H-bond energies monotonically decreases from the neighbor backbone atoms C & N to the farthest atoms from Oxygen in the amino acid side chains.

It is however notable that once WCN is recalculated using the geometric center of the side chains as the representative coordinates of individual sites, the quantity WCN still outperforms all other structural quantities, including those derived from Voronoi tessellation, in explaining site-specific sequence variability (Figure 4.7 & 4.8). The better performance of WCN compared to local packing density as measured from Voronoi cell volume and area may not be surprising, since WCN by its definition takes into account the potential long-range interactions among amino acids in different regions of protein. Indeed, the fractal dimension of proteins very much resembles that of lattice percolation models [114] and similarly the random packing of hard spheres near percolation threshold [60,67]. To expand on this, define the average maximum extent of a protein as,

$$R_m = \frac{1}{2d} \sum_{i=1}^{d} \big(x_{i,max} - x_{i,min}\big), \tag{4.3}$$

in which $d = 3$ is the dimension of the Euclidean space $x_{i,max} - x_{i,min}$ is the maximum physical extent of the protein, as represented by the geometric center of the side-chain coordinates, in each of the three spatial dimensions. Alternatively, the radius of gyration of a protein of length N can be defined as (similar to that of a finite size cluster: Egn. 45a, Sec. 3.2, in [114]),

$$R_g = \sqrt{\sum_{i=1}^{N} \frac{|\vec{r}_i - \vec{r}_0|^2}{N}}, \tag{4.4}$$

where,

$$\vec{r}_0 = \sum_{i=1}^{N} \frac{\vec{r}_i}{N}, \tag{4.5}$$

is the geometric center of the protein, and r_i is the position of the geometric center of the side-chain of the i^{th} amino acid in protein. This definition is such that the kinetic energy and the angular momentum of the protein about the $\vec{r}_0$ is equivalent to the kinetic energy and the angular momentum of all amino acids residing on a ring of radius R_g centered at $\vec{r}_0$. Figure 4.11 & 4.12 depict the behavior of the maximum extent (R_m) versus protein volume (V) and the radius of gyration (R_g) versus the protein length (N), respectively.

The protein volumes and surface areas calculated using *3V* software [119]. A symmetric linear fit (i.e., Deming regression) to the plot of $\log R_m$ vs. $\log V$ and $\log R_g$ vs. $\log N$ results in regression slopes of $D \simeq 2.47 \pm 0.06$ & $D \simeq 2.60 \pm 0.08$ respectively. The observed exponents are very similar to those of the scaling relation,

$$V \propto R^D, \qquad (p \simeq p_c) \tag{4.6}$$

with $D \simeq 2.51 \pm 0.01$ in numerical simulations of hard-sphere packing [67] in three dimensions near percolation threshold ($p \simeq p_c$) or the exponents derived

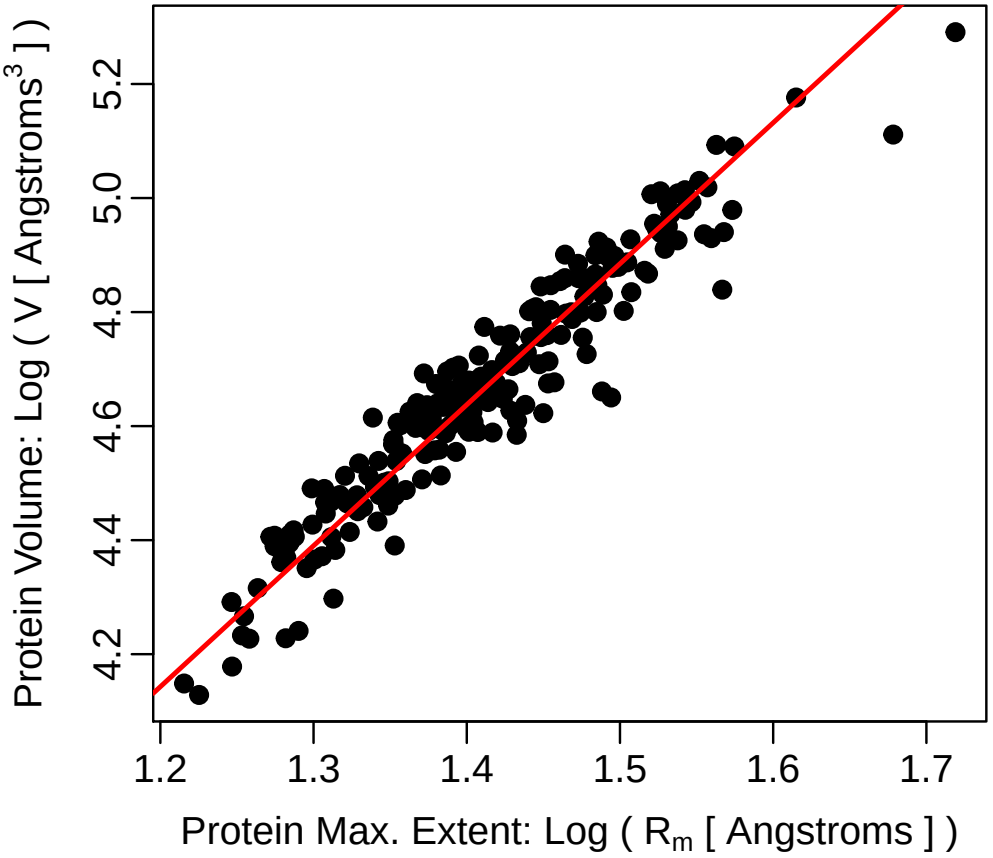

Figure 4.11: The scaling behavior of protein maximum extent as defined by Eqn. 4.3 with protein volume for 209 monomeric enzymes in the dataset. The red line is the linear Deming regression fit to logarithms of the two variables with a slope of $D \simeq 2.47 \pm 0.06$.

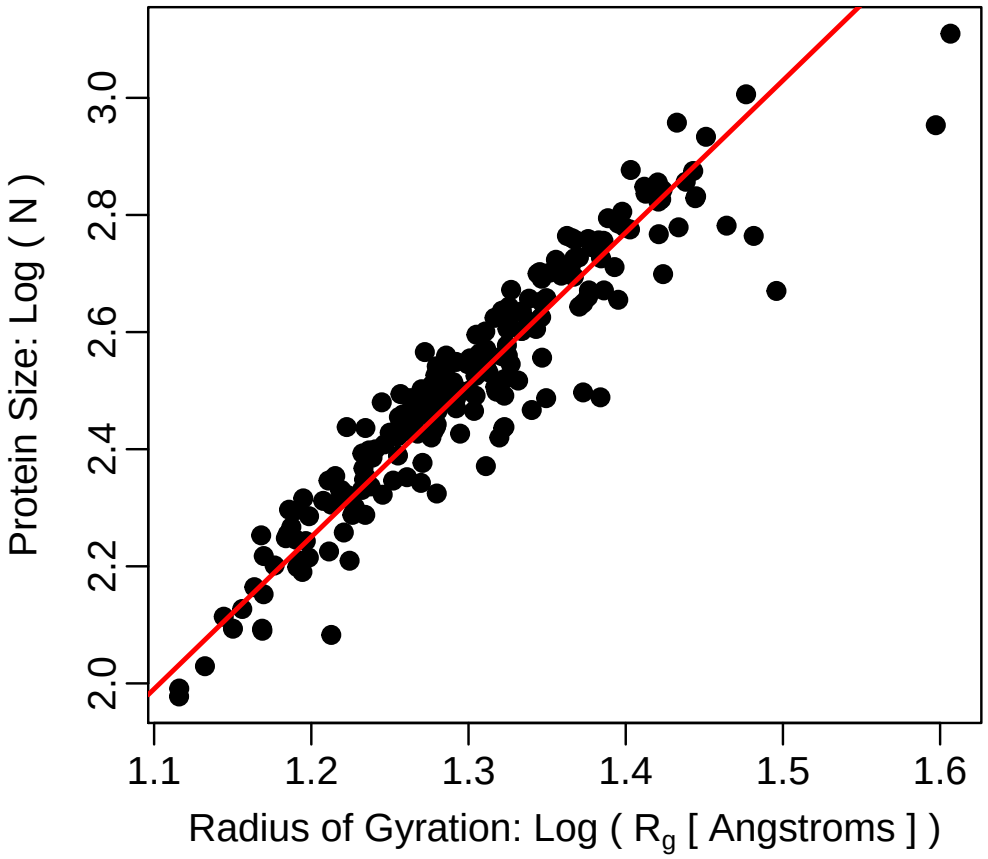

Figure 4.12: The scaling behavior of protein's radius as defined by Eqn. 4.4 with protein length for 209 monomeric enzymes in the dataset. The mean & median length of the proteins are 362 & 315 respectively. The red line is the linear Deming regression fit to logarithms of the two variables with a slope of $D \simeq 2.60 \pm 0.08$.

from lattice models $D \simeq 2.54 \pm 0.05$ [1] & $D \simeq 2.5$ [114]. It is notable that far from percolation threshold (i.e., $p \to 0$) in three dimensions, $D = 2$ [85], while above the threshold ($p > p_c$), $D = d = 3$, in 3D space.

4.5.1 Side-Chain vs. C_α B Factors in Representing Site-Specific Fluctuations

The observed improvements in correlations of average side-chain B factor (vs. C_α B factor) with other structural properties also merit further attention. It was discussed in Section 4.4 and depicted in the plots of Figure 4.6 that in general, as one moves from the B factors of atoms in the backbone of amino acid to the B factor of side-chain atoms, the correlations of B factor with other site-specific structural and sequence properties improve. In particular, the use of average side-chain B factor turned out to result in the highest correlation strengths with other site-specific properties, implying that this average B factor is likely the best representation of the overall amino acid fluctuations and flexibility in a given site in protein. The definition of B factor and its derivation from Debye-Waller factor has been already discussed in Chapter 2, Eqns. 2.8–2.10.

The mean-square-displacement $\langle u^2 \rangle$ in Eqn. 2.10 can be decomposed into four contributing components [27],

$$\langle u^2 \rangle = \langle u^2 \rangle_c + \langle u^2 \rangle_d + \langle u^2 \rangle_{ld} + \langle u^2 \rangle_v, \tag{4.7}$$

in which subscripts c, d, ld, v refer to fluctuations due to conformational sub-

states, diffusion, lattice disorder, and thermal vibrations respectively. The second term $\langle u^2 \rangle_d$ is generally negligible and can be ignored in Eqn. 4.8. Of particular interest to this study is the first term, which is also typically the major contributor to the overall value of the atomic B factor, specially in high-resolution X-ray crystallography of proteins. This term represents the positional displacements of the atom of interest together with other atoms in the amino acid between many different conformational substates of the protein, with the transition probability between the substates governed by the Boltzmann distribution. Compared to atomic coordinates, there are comparatively fewer restraints on the atomic B factors during X-ray crystallography refinement process, and thus in this regard B factor is generally considered as the *error sinks* for static and dynamic disorder and various kinds of model errors in the refinement process [91]. The noise and model uncertainty contributions to the atomic B factors in particular increase with decreasing the resolution of the X-ray crystallography. Better resolution in general corresponds to lower average B factors for the entire structure of the protein [91].

Although the extraction of conformational fluctuations from noise in B factors seems a daunting task [91], the effects of noise, model error and uncertainties due to limited X-ray crystallography resolution can be minimized by averaging B factors over the entire amino acid in a given site: To expand on this, consider the contribution of conformational fluctuations between different substates to be approximately the same for all atoms in the amino acid. The conformational fluctuations can be regarded as the collective motion of all

atoms in the amino acid, on top of which there are noise fluctuations in each of the atoms. These collective motions are the type of fluctuations in B factors that are expected to reflect the biologically relevant and important factors for the proper functioning of the protein. The stochastic noise in the fluctuations is often assumed to have an isotropic Gaussian origin. Therefore, averaging over the atomic B factors in each individual amino acid essentially results in higher Signal-to-Noise Ratio (SNR) in the measurement of the amino acid conformational fluctuations. Figure 4.13 illustrates how this averaging over all atomic B factors increases the SNR in measuring the fluctuations due to conformational substate transitions of the amino acid.

To expand further on this, a simple argument may be given to explain the observed strongly-positive approximately-linear correlation between the two parameters in the plot of Figure 4.13. The contributions to the atomic B factor values of the i^{th} atom in the amino acid in the j^{th} site in a given protein can be assumed to originate from two major sources: conformational substates and stochastic noise due to model uncertainties in refinement process and limited resolution of the X-ray crystallography,

$$\langle u^2 \rangle_{ij} = \langle u^2 \rangle_{\text{substates},ij} + \langle u^2 \rangle_{\text{noise},ij}. \tag{4.8}$$

For simplicity and without loss of generality, one can assume that the contribution of fluctuations due to conformational substate transitions is approximately the same for all atomic B factors in a given amino acid residing

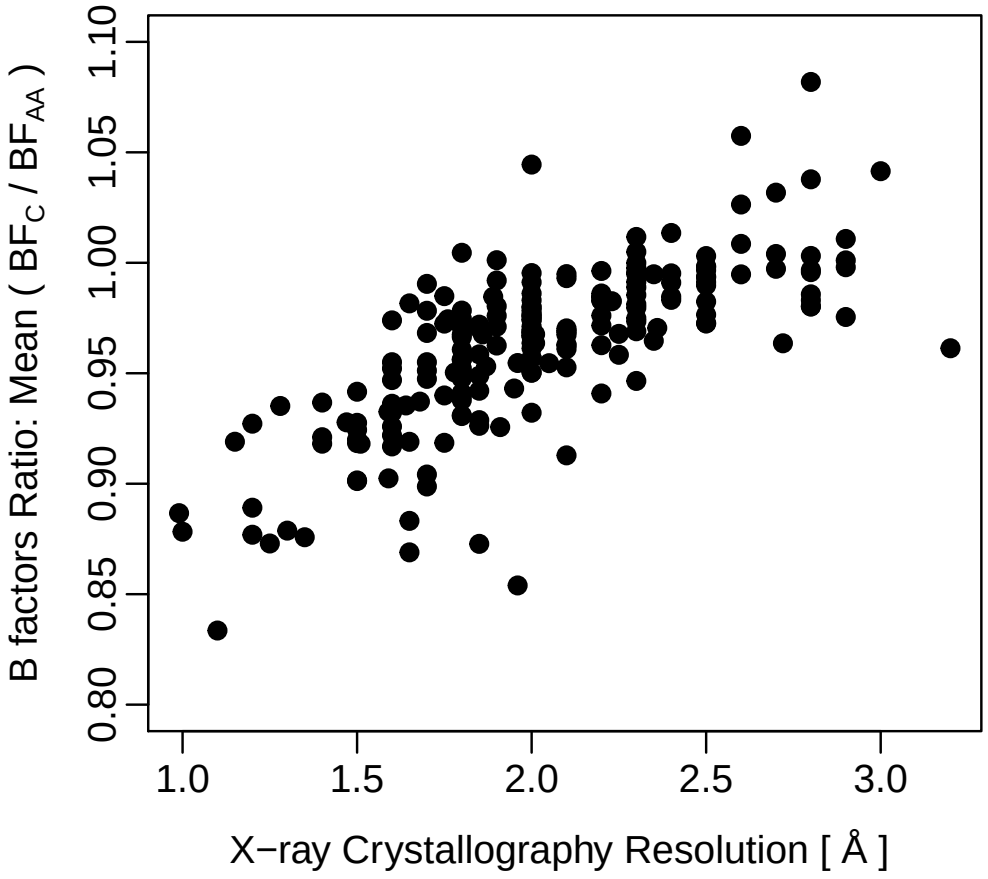

Figure 4.13: An illustration of the strong positive correlation of X-ray crystallography resolution with the ratio of the backbone C atomic B factor to the average amino acid B factor ($\mathrm{BF}_C/\mathrm{BF}_{AA}$), averaged over all sites in individual proteins, highlighting the significant contributions of noise and model errors to atomic B factor values. The Spearman's correlation coefficient between the two quantities is $\rho \sim 0.76$. No significant correlation would be expected in the absence of noise due to limited resolution of the X-ray crystallography of proteins. Each filled circle in the plot represents one protein in the dataset of 209 enzymes used in this work.

the j^{th} site. In other words, the term $\langle u^2 \rangle_{\text{substates},ij}$ in the above equation has almost the same value $\langle u^2 \rangle_{\text{substates},j}$ for all atoms in the amino acid in the j^{th} site in protein. Thus, the average B factor for the entire amino acid molecule of size N_j atoms would be,

$$\begin{aligned}
\langle u^2 \rangle_j &= \frac{1}{N_j} \sum_{i=1}^{N_j} \langle u^2 \rangle_{\text{substates},ij} + \langle u^2 \rangle_{\text{noise},ij} \\
&= \langle u^2 \rangle_{\text{substates},j} + \frac{1}{N_j} \sum_{i=1}^{N_j} \langle u^2 \rangle_{\text{noise},ij} \\
&= \langle u^2 \rangle_{\text{substates},j} + \frac{1}{N_j} \sum_{i=1}^{N_j} \mu_{\text{noise},j} \qquad (4.9)
\end{aligned}$$

in which $\mu_{\text{noise},j}$ is the average noise in the j^{th} amino acid. The ratio of the B factor of the ij^{th} atom to the average B factor of the j^{th} site in protein can be approximated as,

$$\begin{aligned}
\frac{\langle u^2 \rangle_{ij}}{\langle u^2 \rangle_j} &\simeq \frac{\langle u^2 \rangle_{substates,j} + \langle u^2 \rangle_{noise,ij}}{\langle u^2 \rangle_{substates,j} + \mu_{noise,j}} \qquad (4.10) \\
&= \frac{1}{1 + \mu_{noise,j} / \langle u^2 \rangle_{\text{substates},j}} \\
&+ \left(\frac{\langle u^2 \rangle_{\text{noise},ij}}{\langle u^2 \rangle_{\text{substates},j}} \right) \frac{1}{1 + \mu_{noise,j} / \langle u^2 \rangle_{\text{substates},j}} \qquad (4.11) \\
&\simeq 1 - \frac{\mu_{noise,j}}{\langle u^2 \rangle_{\text{substates},j}}, \qquad (4.12)
\end{aligned}$$

where from line 4.11 to 4.12, an assumption was made that the second term in line 4.11 could be neglected compared to the first term and that the noise compared to conformational fluctuation is small, that is, $\mu_{noise,j} / \langle u^2 \rangle_{\text{substates},j} < 1$

(an error of $0.2\mathring{A}$ corresponds approximately to $1\mathring{A}$ increase in B factor [91]). Knowing that the average noise across different amino acids is approximately the same [27], that is $\mu_{noise,j} \sim \mu_{noise}$, and that the noise due to X-ray crystallography almost negatively linearly correlates with crystallography resolution in the range $\sim 1-3$ $[\mathring{A}]$ [91], that is $\mu_{noise} \propto -$resolution, a positive approximately-linear relationship between the average of the B factor ratios over the entire amino acids in the protein structure and the X-ray crystallography resolution would be obtained,

$$\frac{\mathrm{BF}_C}{\mathrm{BF}_{AA}} = \frac{1}{L}\sum_{j=1}^{L}\frac{\langle u^2\rangle_{ij}}{\langle u^2\rangle_j} \tag{4.13}$$

$$\propto -\mu_{noise}\sum_{j=1}^{L}\frac{1}{\langle u^2\rangle_{substates,j}} \tag{4.14}$$

$$\propto \text{resolution} \tag{4.15}$$

in which L represents the length of the protein sequence. The summation term in line 4.14 would not influence this linear relationship, causing only scatter in the relation, so long as the length of the protein not does impose limitations on the resolution of X-ray crystallography of proteins. In general, however this may not be the case. For the sample of 209 proteins considered here, there exists indeed a weak Spearman's correlation coefficient of $\rho \sim 0.2$ between protein length (L) and resolution. Figure 4.13 illustrates the relationship between the average B factors ratio and the resolution in the dataset, using atom C in the backbone of all amino acids in proteins representing the i^{th} atom in the

notation of Eqn. 4.13. It is also notable that the the atomic fluctuations due to conformational substates may not be exactly the same for all atoms in an amino acid in a given site in protein. Indeed, one may expect the conformational fluctuations in the backbone atoms would be less significant compared to conformational fluctuations of side-chain atoms.

Although averaging B factor over the entire amino acid atoms would reduce the noise further than averaging over side-chain atoms, the functionally important conformational fluctuations that are better captured by the side-chain atomic B factors would compensate for the increase in the noise, such that overall, the B factors averaged over side-chain atoms results in slightly better correlations with sequence variability and other relevant structural characteristics depicted in Figure 4.6.

4.5.2 Long-Range Amino Acid Interactions Effects on Sequence Evolution

The Weighted Contact Number as defined by Eqn. 4.1 was shown to outperform all other site-specific structural quantities in predicting sequence variability, in particular, by about 8% compared to the second best predictor of sequence evolutionary rates (the Voronoi cell area) as depicted in Figure 4.7. This correlation strength improvement, though minor, reflects the importance of long-range interactions among amino acids. The question however, remains as to what weighting function best represents the long-range residue-residue interactions in proteins. The power-law kernel as used in Eqn. 4.1 has been

almost unanimously used and adopted by all researches ever since its intrduction [43, 64, 104, 126]. There has been however, no physical explanation for the power-law kernel as the optimal representation of long-range interactions in proteins. Furthermore, the widely used value $\alpha = 2$ for exponent of the power-law kernel in the definition of WCN (Eqn. 4.1) may not necessarily correspond to the optimal value.

Here in order to explore the effects of free adjustable parameter α of the power-law kernel in WCN definition (Eqn. 4.1) in modelling the long-range interactions, the Spearman's correlation coefficient ρ between WCN and four other structure and sequence characteristics were calculated for all 209 proteins in the dataset, for a wide range of exponent values ($-30 < \alpha < 30$). The site-specific characteristics considered include the evolutionary rates ($r4sJC$), sequence entropy, $\Delta\Delta G$ rate, and average side-chain B factors. The results are illustrated in Figure 4.14. The exponent values resulting in the highest correlation strengths together with the median values of the correlation strength distribution for all proteins in the dataset are tabulated in Table 4.1.

Alternatively, once could consider other weighting functions in the definition of the WCN. Here for comparison, the following definition was considered and studied,

$$\mathrm{wcn}_i = \sum_{j=1}^{L} \exp\left[-\left(\frac{r_{ij}}{s}\right)^{\gamma} \right] \quad ; \qquad j \neq i, \tag{4.16}$$

in which L represents the length of the protein sequence, and s is a scale

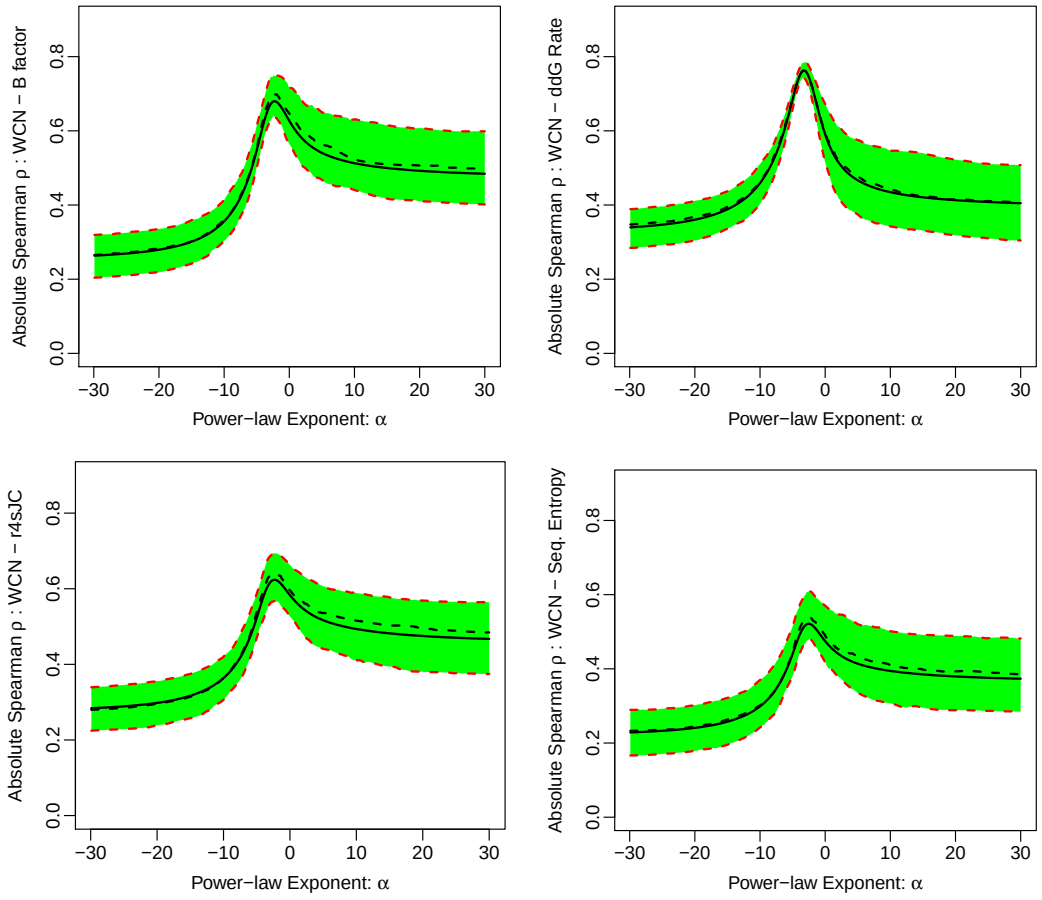

Figure 4.14: The average absolute Spearman's correlation strengths of the Weighted Contact Number with power-law kernel as given by Eqn. 4.1 for different values of the free parameter of the kernel α. The solid black line represents the mean correlation strength in the entire dataset of 209 proteins, and the dashed black line indicates the median of the distribution. The green-shaded region together with the two read dashed lines represent the 25% & 75% quartiles of the correlation strength distribution. Note that for $\alpha > 0$ the sign of the correlation strength ρ is the opposite of the sign of ρ for $\alpha < 0$. In addition ρ is undefined at $\alpha = 0$ and not shown in this plot. The parameter values at which the Spearman's correlation coefficient reaches the maximum over the entire dataset are given in Table 4.1.

Table 4.1: Best free parameters of different definitions of WCN (using four different weighting functions: power-law, exponential, Gaussian, and cutoff distance) that result in the strongest median Spearman's correlation (ρ) of WCN with four site-specific quantities (average side-chain B factor, evolutionary rates (r4sJC), sequence entropy, and $\Delta\Delta G$ rate) for the entire dataset of 209 proteins. The corresponding median correlation coefficients (ρ) are reported inside parenthesis next to each parameter value in the table.

Correlation	Best Performing Free Parameter of the Kernel			
	Power-law α	Exponential λ [$\mathring{A}$]	Gaussian σ [$\mathring{A}$]	Cutoff r [$\mathring{A}$]
WCN–B factor	-2.2 (0.70)	3.8 (0.70)	9.6 (0.70)	13.0 (0.69)
WCN–r4sJC	-2.3 (0.64)	3.4 (0.64)	9.8 (0.63)	14.8 (0.62)
WCN–Seq. Entropy	-2.2 (0.54)	3.4 (0.53)	8.0 (0.53)	15.2 (0.52)
WCN–$\Delta\Delta G$ Rate	-3.4 (0.76)	2.0 (0.78)	5.4 (0.79)	9.8 (0.75)

parameter. For $\gamma = 2$ & $\gamma = 1$ the weighting function corresponds to the Gaussian and exponential kernels, respectively with scale parameters σ & λ. For these two specific cases, the behavior of WCN correlation with other site-specific structural properties are also depicted in the plots of Figures 4.15 & 4.16. For comparison, similar plots of the same correlation strengths were also made for the original simple definition of Contact Number (Figure 4.17), in which WCN for the i^{th} site represents the number of amino acids in a spherical neighborhood of radius R_C (i.e., the cutoff distance) around the site of interest.

Contrary to the arguments of Yang et a. (2009) [126], the power-law kernel with an exact exponent $\alpha = -2$ appears to represents neither the best exponent, nor the best choice of the weighting function, even among the few example kernels considered here. Corroborating [41], it is likely that there might indeed not exist a universal long-range interaction model for the energy

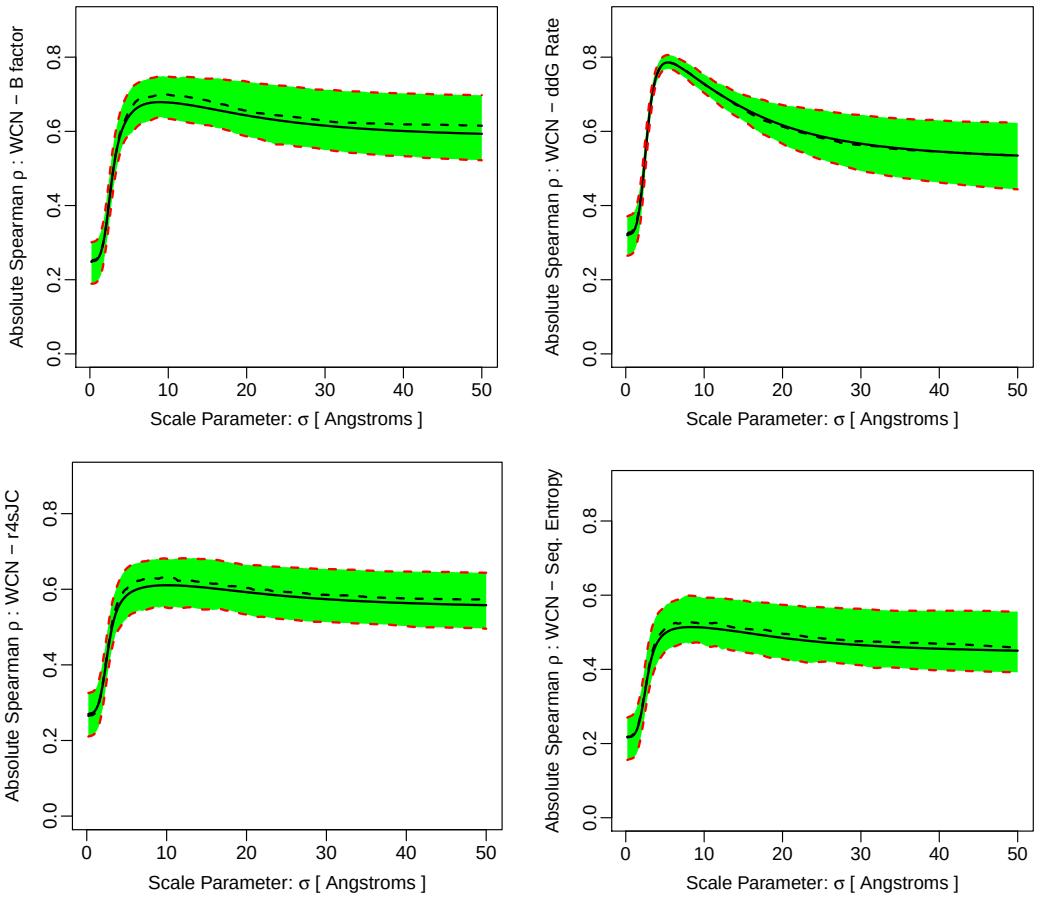

Figure 4.15: The average absolute Spearman's correlation strengths of the Weighted Contact Number with Gaussian kernel as defined by Eqn. 4.16 for different values of the free parameter of the kernel σ. The solid black line represents the mean correlation strength in the entire dataset of 209 proteins, and the dashed black line indicates the median of the distribution. The green-shaded region together with the two read dashed lines represent the 25% & 75% quartiles of the correlation strength distribution. The parameter values at which the Spearman's correlation coefficient reaches the maximum over the entire dataset are given in Table 4.1.

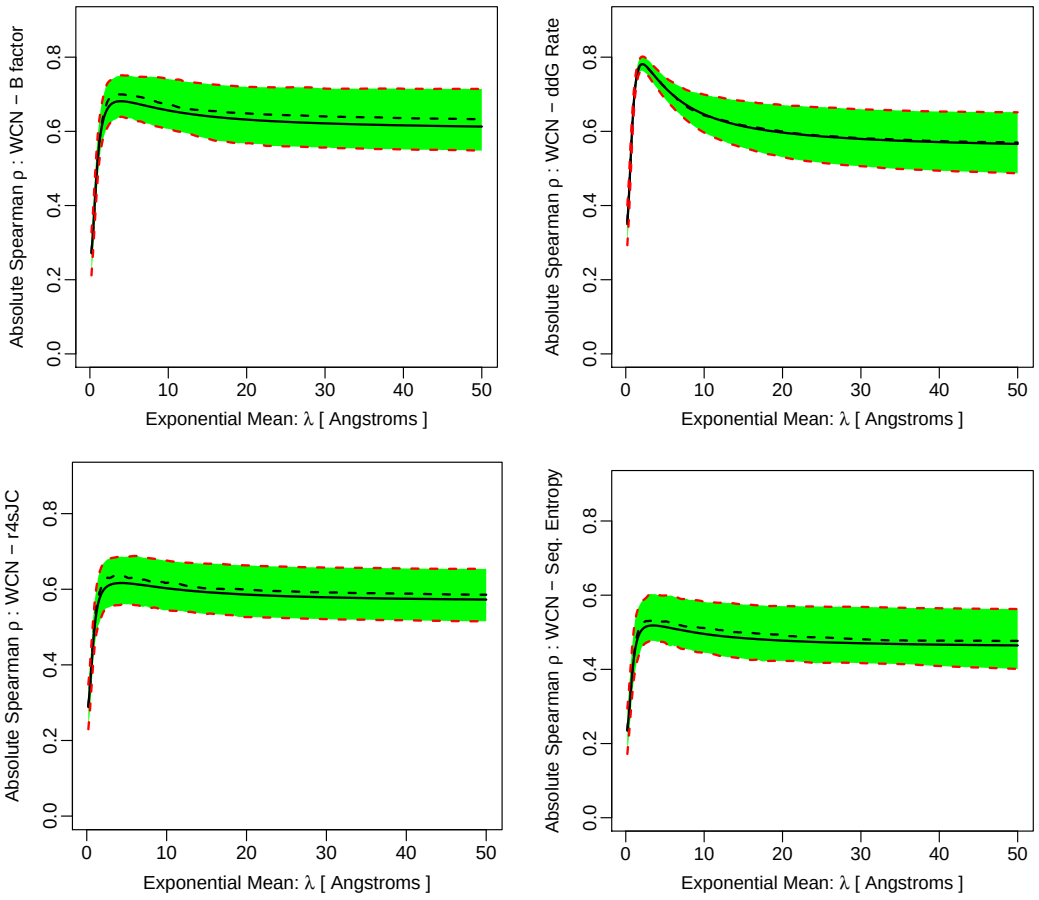

Figure 4.16: The average absolute Spearman's correlation strengths of the Weighted Contact Number with exponential kernel as defined by Eqn 4.16 for different values of the free parameter of the kernel (the exponential mean λ). The solid black line represents the mean correlation strength in the entire dataset of 209 proteins, and the dashed black line indicates the median of the distribution. The green-shaded region together with the two read dashed lines represent the 25% & 75% quartiles of the correlation strength distribution. The parameter values at which the Spearman's correlation coefficient reaches the maximum over the entire dataset are given in Table 4.1.

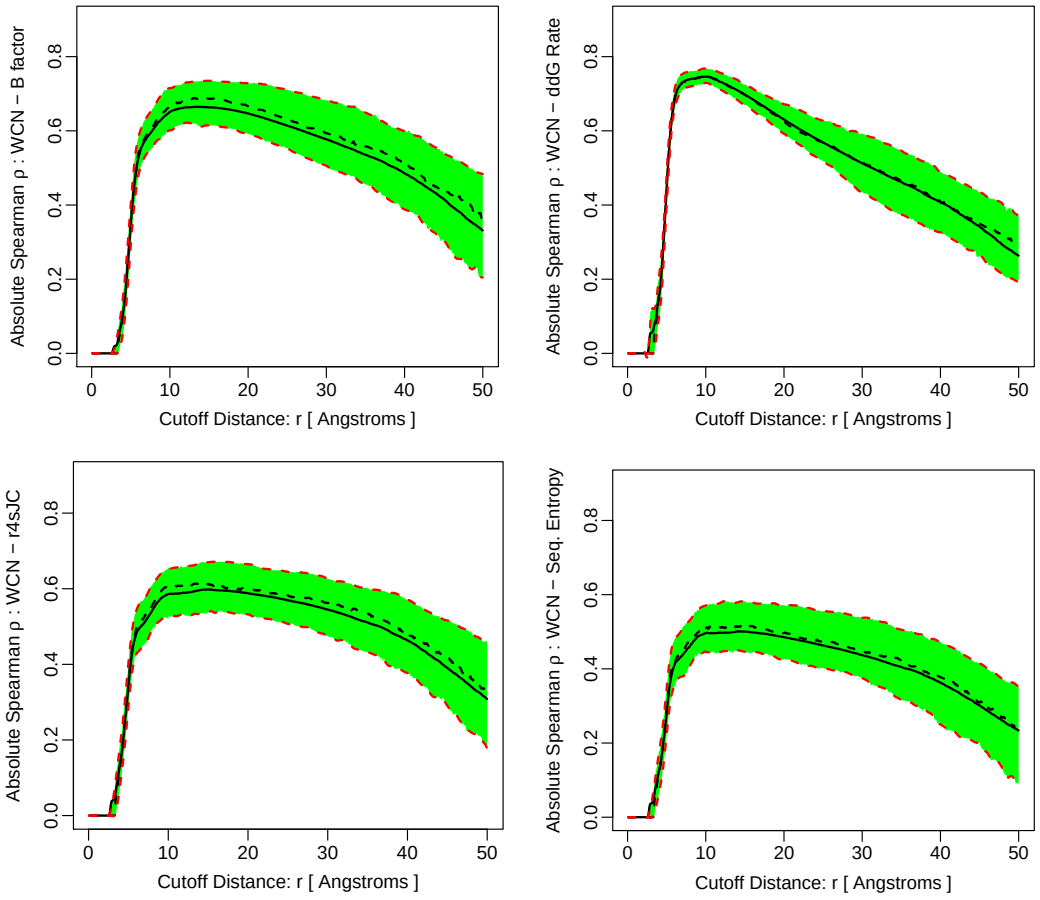

Figure 4.17: The average absolute Spearman's correlation strengths of the Weighted Contact Number with hard-sphere cutoff kernel for different values of the free parameter of the kernel (R_C). This definition of WCN measures of the number of amino acids within a spherical neighborhood of radius R_C around a given amino acid in the site of interest. The solid black line represents the mean correlation strength in the entire dataset of 209 proteins, and the dashed black line indicates the median of the distribution. The green-shaded region together with the two read dashed lines represent the 25% & 75% quartiles of the correlation strength distribution. The parameter values at which the Spearman's correlation coefficient reaches the maximum over the entire dataset are given in Table 4.1.

landscape of all proteins. Alternatively, it may be that the overall contributions of long-range (compared to short-range) interactions in protein dynamics are too small, such that the existing uncertainties in B factors, sequence variability and $\Delta\Delta G$ rates do not allow the precise determination of the functional form of long-range interactions. Alternatively, it may be possible to model the long-range interactions between amino acids in proteins by virtual phonon exchange [83].

For the case of power-law kernel, the plots of Figure 4.14 show that the correlation strengths do not sharply vanish for $\alpha > 0$. The existence of nonzero correlations in the positive range of α has obviously no physical interpretation, otherwise it implies that the farther the two amino acids are from each other, the stronger they interact. Instead, a geometrical argument may be able to explain the observed nonzero correlation in the positive exponent range: In general, amino acid sites on the surfaces of proteins tend to evolve faster than those buried in the core (e.g., Figure 4.10). Assuming a globular shape for the majority of proteins, a power-law WCN with positive exponent would result in large WCN values for sites that are on the surface of the protein, whereas site that are buried close to the center of mass of the protein would have lower WCN. Therefore WCN with positive exponent, is simply a proxy measure of the closeness of the sites to the geometrical center of the protein.

Chapter 5

Identifying the Structural and Evolutionary Modulators of the Strength of Sequence-Structure Relations

5.1 Introduction

Patterns of amino acid sequence variation are known to be influenced by the function of proteins. The general consensus, based on the flurry of research done over the past several decades, is that the amino-acid sequence determines the 3D structure of proteins, known as the native conformation. This sequence-structure relation, however, does not necessitate a unique one-to-one mapping of sequence and the functionality of the protein. According to stability threshold model of proteins [6], some amino acid substitutions at specific sites may be tolerated, if the new amino acid does not significantly change the energy landscape of the protein and therefore, its functional *native* conformation. Indeed, it has been already shown in Chapters 3 & 4 that site-specific structural properties can explain the general patterns of sequence variability in proteins. One of the earliest discovered examples of such relations, is the correlation of the site-specific Relative Solvent Accessibility (RSA) with measures of sequence variability such as sequence entropy and different measures of site-specific evolutionary rates. Amino acid residues that are buried in the

core of proteins tend to be more evolutionary conserved than exposed residues close to the surface of the protein.

Other structural properties were also identified and proposed in previous chapters to influence or explain the site-specific evolutionary variations of proteins. Among the simplest properties is the residue *contact number (CN)*, a measure of local density of the protein defined as the number of amino acids within a spherical neighborhood of a specific residue of interest. Variants of this quantity that attempt to eliminate the free-parameter (i.e., the radius of the spherical neighborhood) in the definition of CN were shown correlate have more explanatory power about sequence evolutionary rates (e.g., Figures 4.7 & 4.8). Other quantities that were shown to correlate strongly with sequence variability measures included the $\Delta\Delta G$ rate, atomic B factors and Voronoi cell characteristics, in particular, the cell volume and area.

Although the majority of proteins exhibit some degree of correlation and association between sequence variation and structural properties, the strength of these correlations vary widely among different structures. As illustrated in Figures 4.7 & 4.8 based on a dataset of 209 monomeric enzymes, a wide range of correlation strengths between sequence variability with site-specific structural characteristics exist. Furthermore, the strengths of sequence-structure relations also tend to correlate strongly with each other, depicted in the plots of Figures 5.1 & 5.2, implying that for a given protein, the correlation strength of a specific structural property with evolutionary rates can serve as a proxy measure of the correlation strength of other structural

properties with sequence evolutionary rates.

The fact that all relevant structural properties seem to have more or less the same predictive power for sequence variability, implies the existence of one or more structural or evolutionary characteristics of proteins that modulate the strengths of all sequence-structure relations in all proteins. Motivated by these observations, here I present the results of comprehensive effort in search for the potential underlying structural or evolutionary properties of proteins that can explain the wide range of variations seen in correlation strengths of sequence evolutionary rates with different structural properties. Among all properties considered, it is shown that sequence divergence appears to be the primary determinant of the strength of virtually all sequence-structure relations. In addition, proteins with more homogeneous Hydrogen bond (H-bond) energies, corresponding to higher fractions of helical secondary structures and lower fractions of β-sheets generally tend to exhibit the strongest sequence-structure relations. In the following sections, evidence in support of these findings will be presented and their implications will be discussed.

5.2 Materials and Methods

5.2.1 Sequence Data, Alignments and Evolutionary Rates

The results presented in this work are based on the same dataset of 209 monomeric enzymes that were previously used in Chapter 4. The proteins are randomly picked from the Catalytic Site Atlas 2.2.11 [87] with protein sizes in the sample ranging from 95 to 1287, including representatives from all six main

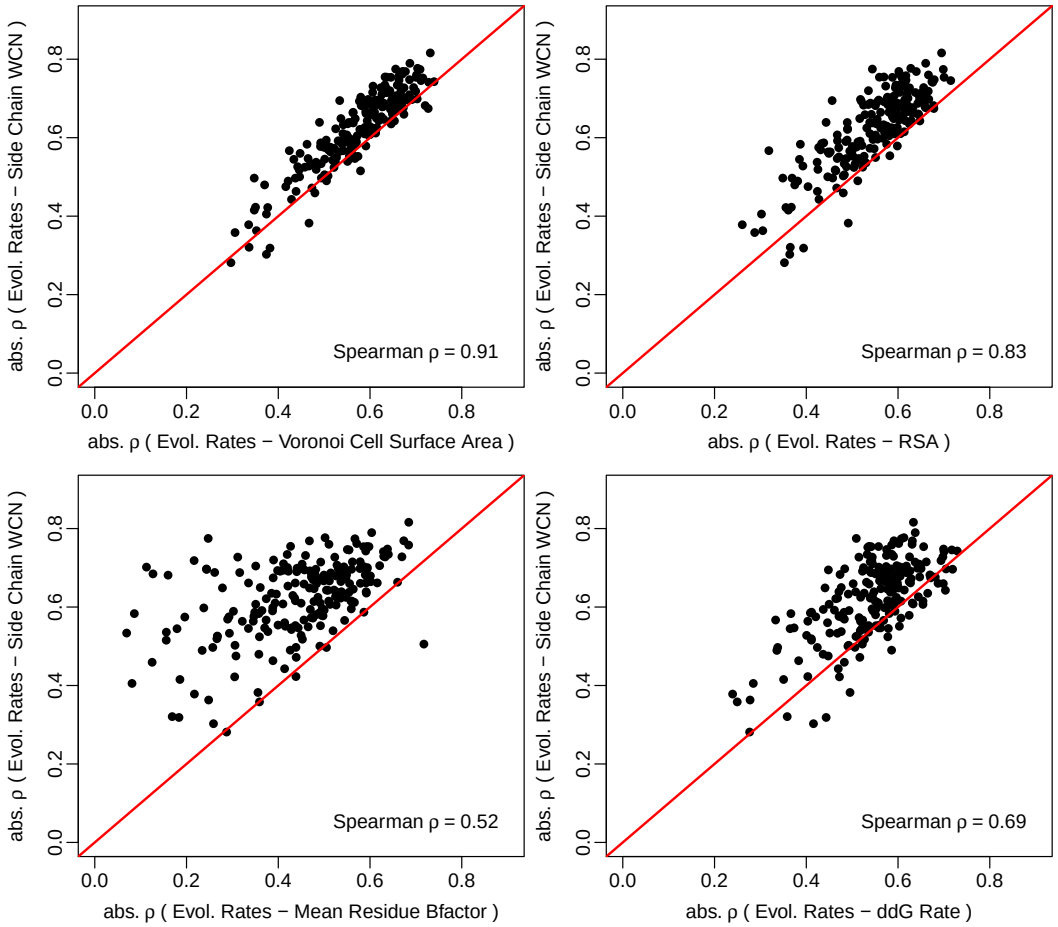

Figure 5.1: A comparison of the strength of the Spearman's correlation strength of sequence evolutionary rates (r4sJC) with side chain Weighted Contact Number (on the vertical axes of plots) vs. correlation strengths of other structural properties with evolutionary rates (on the horizontal axes). Detailed description of the structural properties is given are given Chapters 2 & 4. The red lines in each plot represent equality. It is evident from all plots that for any given protein in dataset, the correlation strength of one structural property is a good proxy measure of the correlation strength of any other structural property with sequence variability measures. The correlation strengths of the two correlation measures on the vertical and horizontal axes are provided on the bottom-right of each plot.

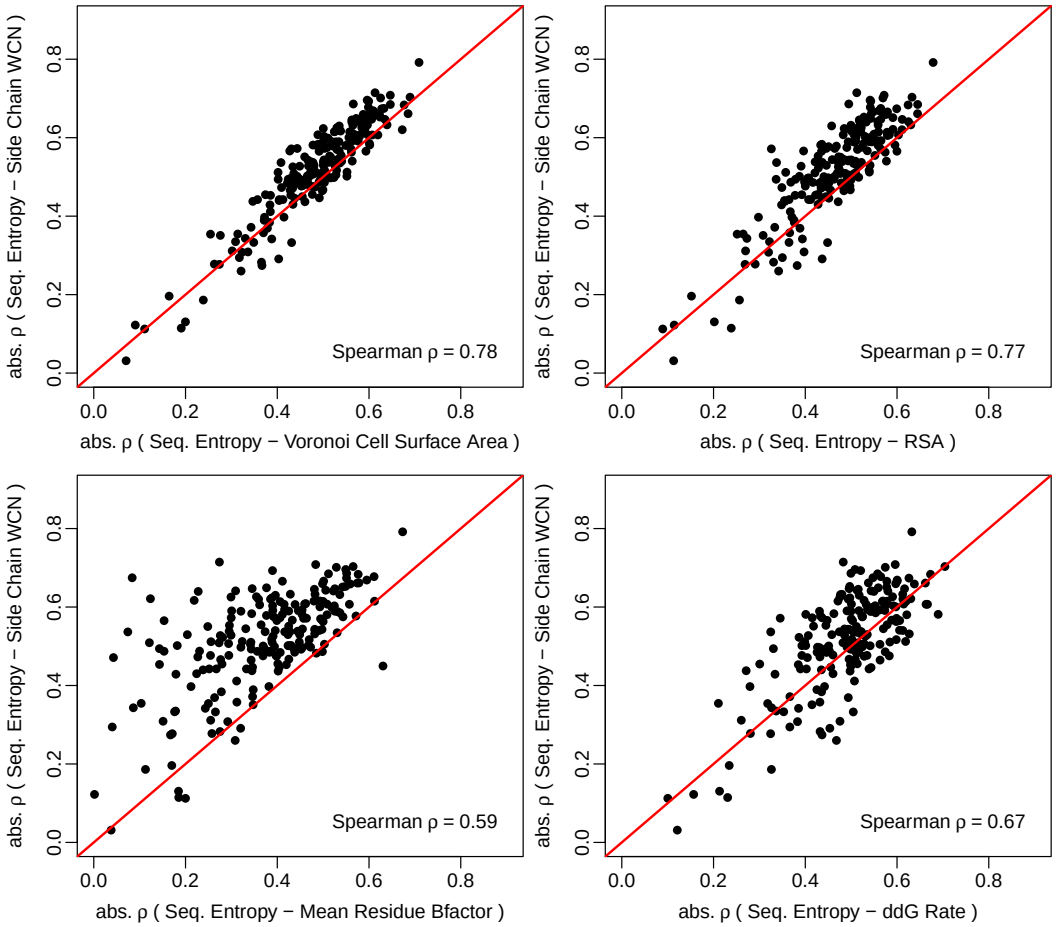

Figure 5.2: A comparison of the strength of the Spearman's correlation strength of sequence entropy with side chain Weighted Contact Number (on the vertical axes of plots) vs. correlation strengths of other structural properties with sequence entropy (on the horizontal axes). Detailed description of the structural properties is given are given Chapters 2 & 4. The red lines in each plot represent equality. It is evident from all plots that for any given protein in dataset, the correlation strength of one structural property is a good proxy measure of the correlation strength of any other structural property with sequence variability measures. The correlation strengths of the two correlation measures on the vertical and horizontal axes are provided on the bottom-right of each plot.

EC functional classes (Webb 1992) and domains of all main SCOP structural classes (Murzin et al. 1995). The process for assessing the evolutionary rates at the amino acid level for each protein, are exactly the same as those described in previous chapters. In addition, the same structural and sequence properties were also collected for the 9 viral proteins that were initially studied in Chapter 3.

5.2.2 Structural Properties

The goal of the presented work is to identify the prominent structural or evolutionary properties of proteins that modulate the strengths of sequence-structure correlations. These modulators potentially represent a unique characteristics of the protein as a whole. In general, the structural and evolutionary properties fall into two major categories. 1. *Residue-level properties*: Site-specific structural or evolutionary characteristics that are defined and calculated for each specific amino acid site in the protein sequence. Prominent examples of the site-specific structural properties were extensively studied in Chapters 2 & 4 and include RSA, WCN, B factor, $\Delta\Delta G$ rate, and Voronoi cell characteristics 2. *PDB-level properties*: structural or evolutionary characteristics that are representative of the protein as a whole. Examples include protein size and compactness, sequence length, structural resolution of the protein in X-ray crystallography. In addition, the distribution of each residue-level property can be summarized by its statistical moments as pdb-level property of the protein. Prime examples include, the mean and variance of WCN, RSA,

sequence entropy, evolutionary rates.

Among other important pdb-level properties that were calculated for all proteins in the dataset is, the protein Contact Order (CO) [86]. This quantity is a measure of the interconnectedness of the amino acids in a protein. For a protein sequence of length L, CO is defined as,

$$\mathrm{CO} = \frac{1}{L \times N} \sum_{i=1}^{L-1} \sum_{j>i}^{L} \Delta S_{ij} \tag{5.1}$$

in which ΔS_{ij} is the distance between the i^{th} & j^{th} amino acids *along the protein sequence*, *iff* the spatial distance between the two amino acids is less than cutoff contact distance r, otherwise $\Delta S_{ij} = 0$. The two amino acids are said to be in contact with each other if $r \lesssim 6$ [$\mathring{A}$].

Other notable protein characteristics that are considered in this work include, the total volume and surface area of the protein and their ratio, giving a measure of the compactness of the protein. In addition, information about the secondary structures of proteins are also extracted using DSSP software [48], such as the total number of residues participating in different types of helices, parallel or anti-parallel beta sheets, or loops and turns. To complete the list of pdb-level structural properties, we also calculate the Spearman correlations between all residue-level structure and sequence properties and include them in the analysis to probe their potential effects on the strength of sequence-structure relations.

A complete description of nearly 372 pdb-level protein properties that

are eventually obtained and calculated for all proteins in the dataset and their detailed presentation here in this work seems impractical. Instead, all data including the list of 209 proteins and their properties together with Python, R and Fortran codes written for data reduction and analysis are publicly available to view and download at `https://github.com/shahmoradi/cordiv`.

5.2.3 Eliminating Degeneracy in Structural Property Definitions

In order to identify the potential determinants of sequence-structure correlations, first a comprehensive search was performed to identify site-specific structural properties that correlate with measures of sequence variability (i.e., sequence entropy & evolutionary rates). There are however degeneracies in the definition of some site-specific characteristics of proteins that need further scrutiny. For example, the quantity WCN is generally calculated from the coordinates of α-carbon atoms in the 3-dimensional structure of proteins as given by Eqn 2.14. There is however no reason to believe this set of atomic coordinates are the best representatives of individual sites in proteins. The same ambiguity also exists as to which set of atomic B factors best represent the site-specific fluctuations in proteins, although the popular choice of residue flexibility is α-carbon atomic B factor [38]. Similar definition degeneracies also exist for the set of coordinates that can be used for Voronoi tessellation of proteins. A popular tool in condensed matter physics, the Voronoi tessellation of a set of points (seeds) is a way of dividing the space into a number of regions such that for each seed there will be a corresponding region consisting of all

points closer to that seed than to any other. These regions are called Voronoi cells. The structure of proteins can be considered as a set of 3D coordinates representing individual sites. Similar to WCN and B factor, there is also ambiguity as to which set of residue atomic coordinates best represent individual sites in proteins for the calculation of Voronoi cells.

As shown in Chapter 4, the structure of proteins appears to be best represented by the geometric center of the side chains of amino acids in individual sites. All other atomic coordinates, in particular those of the backbone atoms tend to contain less information about protein structure. Based on these observations, only variables measured from average side chain properties and coordinates are kept throughout the rest of the analysis and all other similar measures that show only weaker correlations with other site-specific characteristics are omitted. The exclusion of these alternative measures results in a significant reduction in the number of pdb-level variables to be further analysed, from 372 to 165 pdb-level characteristics, without compromising generality and comprehensiveness of the analysis.

5.3 Results

5.3.1 Sequence Divergence as the Main Determinant of Sequence-Structure Relation

In order to identify the potential contributing factors to the strength of sequence–structure correlations, one can first employ one of the simplest nonparametric, yet powerful tests of statistical dependence, that is, the Spear-

man correlation matrix of all pdb-level structure and sequence properties is first constructed. The Spearman' correlation coefficient versus the popular Pearson's correlation measure is chosen, in order to minimize the effects of any nonlinear variable relationships on the strengths of the correlations. The resulting correlation matrix reveals a myriad of pdb-level properties each having a small but nonzero contribution to the strength of the structure-sequence correlations.

A hierarchical clustering of the correlation matrix however, reveals two main independent factors that have the strongest influence on the strengths of sequence-structure correlations (Figure 5.3): 1. The sequence divergence as measured by the variance (or equivalently the standard deviation) of sequence entropy and evolutionary rates (denoted by *sd.seqent* & *sd.r4sJC*) among all sites in each protein structure. It should be however noted that the variance of different sequence variability measures also reflect the ability of the specific sequence variability measure to capture sequence divergence from the sequence alignments. In the following lines, it will be shown that this may be indeed the case with the two sequence variability measures considered in this work. 2. The homogeneity of the hydrogen bond strengths among the back bone atoms of each protein structure, as measured by the variance (or equivalently the standard deviation) of hydrogen bond energies (denoted by *sd.hbe*) among all pdb sites.

A reduced-size of the Spearman's correlation matrix for the most influential factors on the strongest sequence–structure relation (r4sJC – WCN) is

illustrated in Figure 5.4.

For the other weaker sequence–structure relations, (i.e. the correlations of seq.entropy/r4s with Voronoi cell characteristics, RSA, $\Delta\Delta G$ rate, and B factor) the same pdb-level properties are found to contribute the most to the correlation strengths. In general, it is observed that for the weaker sequence-structure correlations, factors that determine the accuracy of the measured residue properties become more influential on the strength of the correlations. In particular, the X-ray crystallographic resolution of the structure and the definition of the $\Delta\Delta G$ rate play dominant roles, having Spearman correlation coefficients of $\rho \sim 0.3$, with the strengths of the corresponding sequence-structure relations.

To ensure the accuracy of the results obtained from the Spearman correlation matrix of the pdb-level properties, we also use multivariate linear regression models, with individual sequence-structure correlations as the sole regressand of the regression models, and the set of pdb-level properties as the explanatory variables. Since the number of explanatory variables is comparable to the number of observations (i.e., the number of pdb structures in the dataset: 209), regularized regression methods [28, 109] were used on the entire dataset, and also on the rank transformation of the dataset in order to minimize the effects of potential nonlinearities in data. Depending on value of the free parameter α, this generalized regression model is a compromise between the *ridge regression* – which attempts to shrink the coefficients of correlated predictors towards each other – and the *lasso regression* – which tends

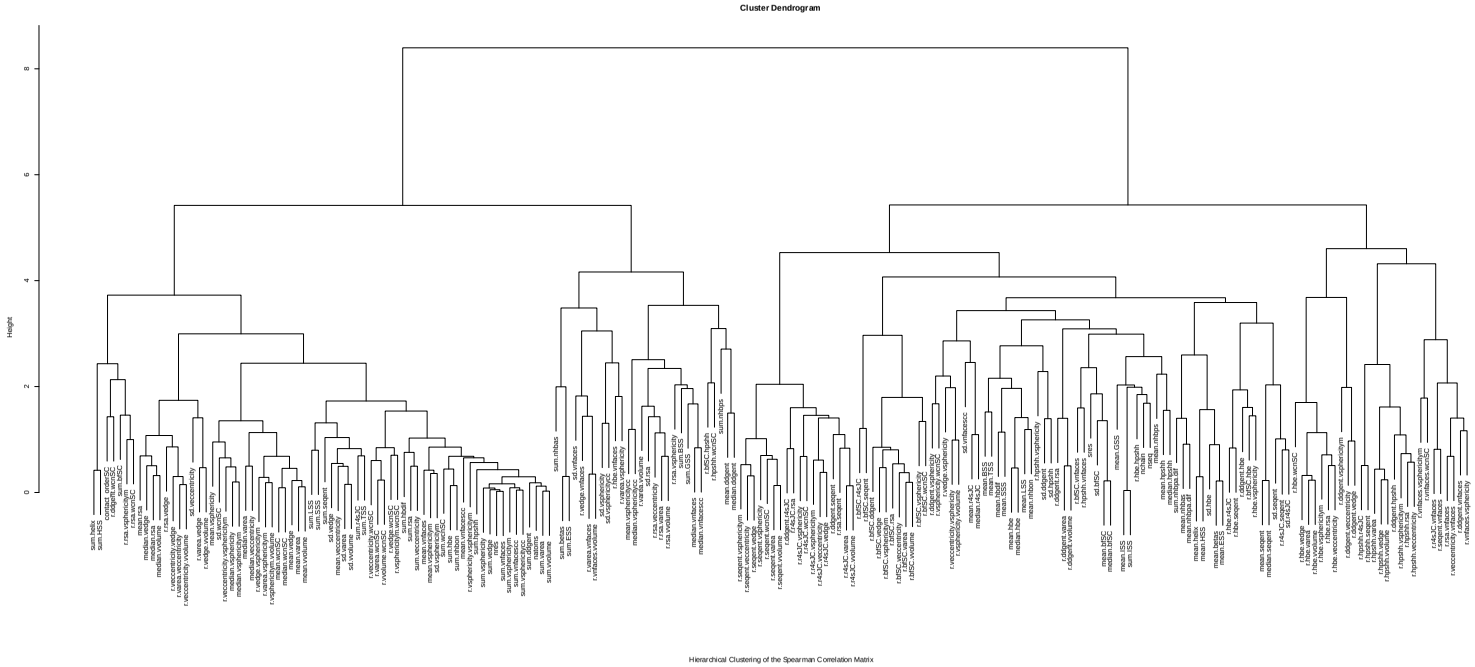

Figure 5.3: Hierarchical clustering diagram of the Spearman's correlation matrix of all pdb-level properties considered in this work, used to identify groups of closely related variables that potentially represent a similar underlying property of proteins. A full size of the diagram and the meaning of each of the variables are available in the permanent online repository of the work (c.f., Section 5.2).

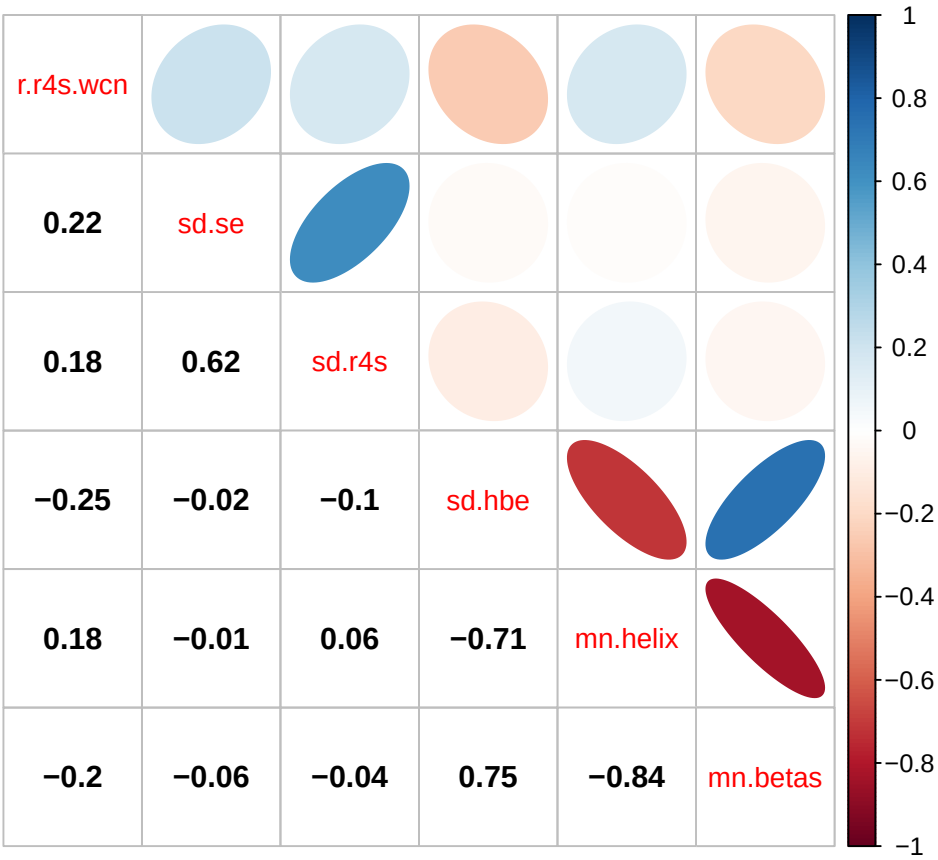

Figure 5.4: **The Spearman correlation matrix for the strongest sequence–structure correlation (denoted by** *r.r4s.wcn*) **and the prominent determinants of the strengths of this relation**. The variables on the diagonal elements of the matrix from top to bottom represent respectively, the strongest sequence–structure relation – i.e., the absolute Spearman's correlation of evolutionary rates (r4s) with side-chain Weighted Contact Number (WCN) – followed by important protein properties that appear to modulate the strength of this relation: variance of sequence entropy (*sd.se*), variance of site-specific evolutionary rates (*sd.r4s*), variance of back-bone hydrogen bond energies (*sd.hbe*), and the fraction of amino acids in helical & $\beta-$sheet secondary structures in the protein (*mn.helix* & *mn.betas* respectively).

to pick one of the correlated predictors and discard the rest. In addition to regularized regression, Principal Component Regression (PCR) methods were also used on the original dataset and its rank transformation. Both regression methods, PCR & regularized, point to similar set of pdb-level properties as the strongest determinants of sequence-structure correlations. The complete list of the regression results are available in the permanent online repository of the work.

5.4 Discussion

Throughout this work, a comprehensive analysis was carried out in search for the main determinants of the strength of sequence-structure relations. Examples of sequence-structure relations include the correlations of sequence entropy and measures of evolutionary rates (such as *r4sJC* used in this work) with measures of local packing density (e.g., *wcnSC*), Relative Solvent Accessibility (RSA) of amino acids, $\Delta\Delta G$ rate (a measure of the stability contribution of sites to protein's native conformation upon random substitutions) and measures of amino acid flexibility in individual sites including B factor and Voronoi cell volume. The majority of these sequence–structure relations were extensively discussed in Chapters 3 & 4.

Overall, it is found that the observed variability in sequence-structure correlation strengths among different proteins is a result of a multitude of structure and sequence factors, each having a small contribution to the variability seen in the strengths of the relations. There are however, two ma-

jor factors from protein sequence and structure that appear to influence all sequence-structure relations similarly and most significantly: The sequence divergence, and the homogeneity of the strength of the backbone hydrogen bonds.

By employing several independent parametric and non-parametric statistics, such as Spearman rank test, regularized regression and Principal Regression methods, sequence divergence is identified as the dominant factor in determining the strength of sequence-structure correlations, capable of explaining $10 - 30\%$ of the observed correlation strengths alone, in both the original and rank-transformed data.

The second most influential protein characteristics appears to be the homogeneity of the backbone hydrogen bond energies. In general, β-sheet secondary structures tend to have less homogeneous H-bond energies than α-helices. It can be therefore concluded that proteins with less fractions of β-sheets and more helical structures tend to exhibit stronger sequence-structure relations. A simple explanation for the observed pattern may be that more homogeneous hydrogen bonds allow the impact of structural characteristics (primarily due to the side-chain interactions) and their variations upon substitution to be more pronounced and visible across the entire structure of protein.

To ensure that the conclusions made in this work remain valid in other protein datasets, the viral protein data from Chapter 3 were also compared to the results from 209 monomeric enzymes presented here. A illustration of the

effects of sequence divergence on the strongest sequence-structure relation in the two datasets of viral and enzymatic proteins is provided in Figure 5.5. The identification of sequence divergence as the main modulator of the strength of sequence-structure relations, highlights the importance of structure in shaping the general patterns of sequence evolution among proteins. In other words, the lack of a significant strong correlation between structural properties and sequence variability of a protein is likely merely a result of inadequate sequence divergence, or inaccurate models used for the calculation of sequence evolutionary rates.

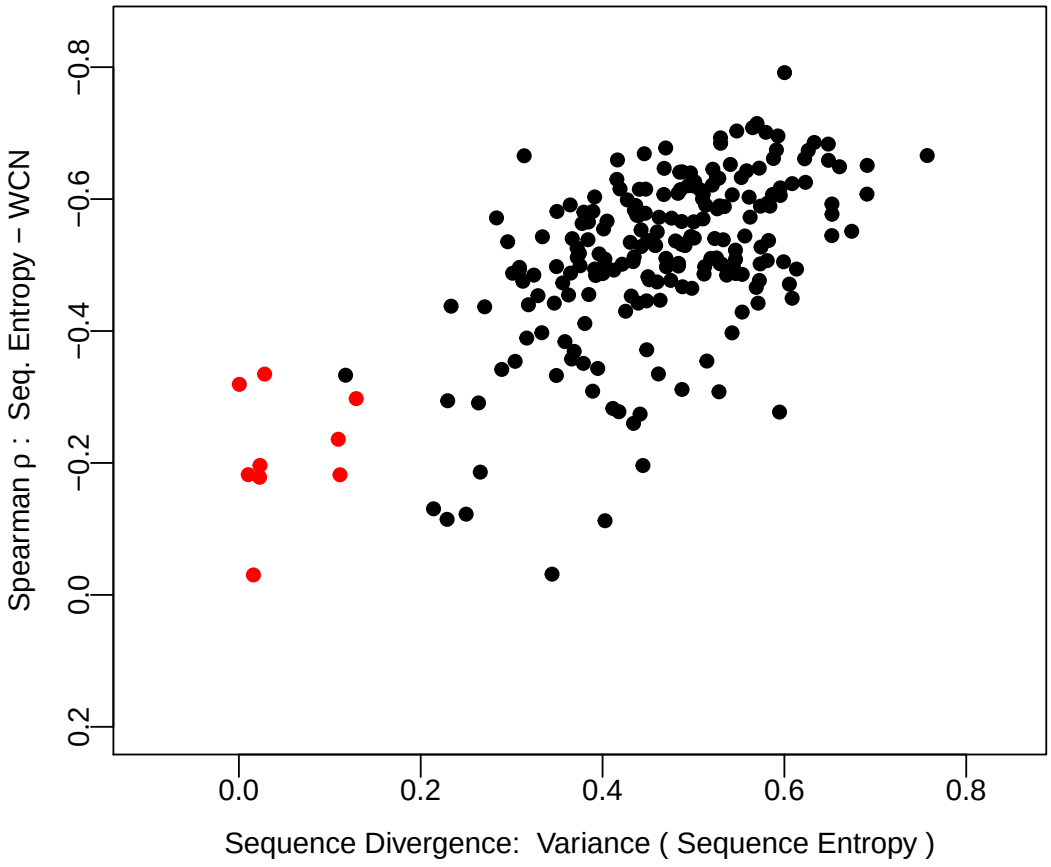

Figure 5.5: **Sequence–structure correlation strength versus sequence divergence.** The plot illustrates the relationship between the strength of a representative sequence–structure correlation (sequence entropy – Weighted Contact Number) and the sequence divergence as measured by the variance of protein sequence entropy. The black circles represent 209 proteins used in this work. For comparison, the red circles represent data from 9 viral proteins taken from Chapter 3 [104].

Chapter 6

Conclusion

Throughout the previous chapters a comprehensive review and analysis of the potential structural determinants of sequence evolution at the amino acid level were presented and discussed. Prime examples of site-specific structural properties include the Relative Solvent Accessibility, measures of Local Packing Density, energetic contributions of individual sites to the stability of the entire protein structure as measured by $\Delta\Delta G$ rate, site flexibility and fluctuations, as measured by atomic B factors, or by quantities derived from Molecular Dynamics simulation such as Root-Mean-Square-Fluctuations (RMSF) and variability in backbone and side-chain dihedral angles (c.f., Chapter 2). Using a dataset of 209 monomeric enzymes, it was shown that the amino acid side-chains play the dominant role in the site's evolutionary variability. It was further shown that the best representation of individual sites in protein's three-dimensional structure is the geometric center of the amino acid side-chains, as opposed to the commonly adopted representation using the C_α backbone atoms.

Among all site-specific structural properties considered, measures of local packing density, in particular the Weighted Contact Number as defined by

Equation 2.14, exhibit the strongest correlations with sequence variability, explaining on average $\sim 41\%$ of sequence variability as measured by site-specific evolutionary rates (e.g., Figures 4.7 & 4.8). It is however important to note that the conventional definition of contact number such as WCN, inherently includes information about long-range side-chain interactions in addition to local packing density of individual sites. In order to segregate the role of long-range amino acid interactions from local packing density in sequence evolution, an alternative method of measuring local packing density was introduced and considered in Chapter 4. The Voronoi partitioning of protein 3-dimensional structure was performed therein on the entire dataset of proteins. It was then shown that the Voronoi cell volume (and similarly Voronoi cell area) provide an ideal measure of the *local* packing density of individual sites in proteins, which unlike WCN exclude the potential effects of long-range interactions in the definition of local packing density. It was then shown that the local packing density (as measured by Voronoi cell volume) can on average explain $\sim 35\%$ of sequence variability represented by site-specific evolutionary rates. By contrast, the amino acid long-range interactions alone as measured from WCN can on average explain $\sim 6\%$ of the observed variability in protein sequences.

It is notable that the strength of sequence-structure relations appears to be primarily modulated by sequence divergence and the quality of sequence alignment and the methodology used for the calculation of evolutionary rates. By contrast, the protein structure appears to have a minor role in modulating sequence-structure relations (c.f., Chapter 5). In other words, the lack of

strong correlation between structural properties of proteins and their sequence variability may simply reflect the lack of sequence divergence or low quality of sequence alignment in a given dataset (e.g., Figure 5.5).

An important question yet remains unanswered in this work which merits further investigation in a future study. The definition of the local packing density, in particular the Weighted Contact Number and the optimal value of its free parameter, has been a matter of extensive debate and discussion over the past years [41,126], although no physical explanation has yet been provided in support of the definitions and the optimal values of the free parameters. The optimal exponent for the power-law definition of WCN has been shown to be ~ -2 [126]. A similar value was also obtained in this work using a different dataset. However, it was also shown in Chapter 4 that other definitions of WCN such as those with exponential and Gaussian kernels (Equation 4.16) can describe the long-range interactions in proteins equally well (Table 4.1). It is therefore perceivable that there might not exist a universal long-range interaction model for the potential energy landscape of all proteins. Alternatively, it may be that the overall contributions of long-range (compared to short-range) interactions in protein dynamics are too small, such that the existing uncertainties in B factors, sequence variability and $\Delta\Delta G$ rates do not allow the precise determination of the functional form of long-range interactions.

Bibliography

[1] Joan Adler, Yigal Meir, Amnon Aharony, and A. B. Harris. Series study of percolation moments in general dimension. *Physical Review B*, 41(13):9183–9206, May 1990.

[2] Haim Ashkenazy, Elana Erez, Eric Martz, Tal Pupko, and Nir Ben-Tal. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Research*, 38(suppl 2):W529–W533, July 2010.

[3] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, 1997.

[4] P. Berens. CircStat: a MATLAB toolbox for circular statistics. *J. Stat. Software*, 31:1–21, 2009.

[5] J. D. Bloom, D. A. Drummond, F. H. Arnold, and C. O. Wilke. Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.*, 23:1751–1761, 2006.

[6] Jesse D. Bloom, D. Allan Drummond, Frances H. Arnold, and Claus O. Wilke. Structural determinants of the rate of protein evolution in

yeast. *Molecular Biology and Evolution*, 23(9):1751–1761, September 2006. PMID: 16782762.

[7] A. J. Bordner and H. D. Mittelmann. A new formulation of protein evolutionary models that account for structural constraints. *Mol. Biol. Evol.*, 31:736–749, 2014.

[8] Erich Bornberg-Bauer and M Mar Alb. Dynamics and adaptive benefits of modular protein evolution. *Current Opinion in Structural Biology*, 23(3):459–466, June 2013.

[9] L. Burger and E. van Nimwegen. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol.*, 6:e1000633, 2010.

[10] R. M. Bush, C. A. Bender, K. Subbarao, N. J. Cox, and W. M. Fitch. Predicting the evolution of human influenza A. *Science*, 286:1921–1925, 1999.

[11] Carlos D. Bustamante, Jeffrey P. Townsend, and Daniel L. Hartl. Solvent accessibility and purifying selection within proteins of escherichia coli and salmonella enterica. *Molecular Biology and Evolution*, 17(2):301–308, February 2000. PMID: 10677853.

[12] Gavin C. Conant and Peter F. Stadler. Solvent Exposure Imparts Similar Selective Pressures across a Range of Yeast Proteins. *Molecular Biology and Evolution*, 26(5):1155–1161, May 2009.

[13] D. R. Cox and H. D. Miller. *The Theory of Stochastic Processes*. CRC Press, February 1977.

[14] A. M. Dean, C. Neuhauser, E. Grenier, and G. B. Golding. The pattern of amino acid replacements in α/β-barrels. *Mol. Biol. Evol.*, 19:1846–1864, 2002.

[15] P. Debye. Interferenz von Rntgenstrahlen und Wrmebewegung. *Annalen der Physik*, 348(1):49–92, January 1913.

[16] N. V. Dokholyan and E. I. Shakhnovich. Understanding hierarchical protein evolution from first principles. *J. Mol. Biol.*, 312:289–307, 2001.

[17] D. A. Drummond, A. Raval, and C. O. Wilke. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.*, 23:327–337, 2006.

[18] J. Echave and F. M. Fernández. A perturbative view of protein structural variation. *Proteins: Structure, Function, and Bioinformatics*, 78:173–180, 2010.

[19] Julian Echave, Eleisha L. Jackson, and Claus O. Wilke. Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. *bioRxiv*, page 009423, September 2014.

[20] R Elber and M Karplus. Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science (New York, N.Y.)*, 235(4786):318–321, January 1987. PMID: 3798113.

[21] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[22] S. J. Fleishman, T. A. Whitehead, D. C. Ekiert, C. Dreyfus, J. E. Corn, E.-M. Strauch, I. A. Wilson, and D. Baker. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, 332:816–821, 2011.

[23] E. A. Franzosa and Y. Xia. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.*, 26:2387–2395, 2009.

[24] E. A. Franzosa and Y. Xia. Independent effects of protein core size and expression on residue-level structure-evolution relationships. *PLoS ONE*, 7:e46602, 2012.

[25] Eric A. Franzosa and Yu Xia. Structural Determinants of Protein Evolution Are Context-Sensitive at the Residue Level. *Molecular Biology and Evolution*, 26(10):2387–2395, October 2009.

[26] Hans Frauenfelder, Guo Chen, Joel Berendzen, Paul W. Fenimore, Heln Jansson, Benjamin H. McMahon, Izabela R. Stroe, Jan Swenson, and

Robert D. Young. A unified model of protein dynamics. *Proceedings of the National Academy of Sciences*, 106(13):5129–5134, March 2009. PMID: 19251640.

[27] Hans Frauenfelder, Gregory A. Petsko, and Demetrius Tsernoglou. Temperature-dependent X-ray diffraction as a probe of protein structural dynamics. *Nature*, 280(5723):558–563, August 1979.

[28] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, 33(1):1–22, 2010.

[29] Mark Gerstein, Erik L. L. Sonnhammer, and Cyrus Chothia. Volume changes in protein evolution. *Journal of Molecular Biology*, 236(4):1067–1078, March 1994.

[30] Ofir Goldenberg, Elana Erez, Guy Nimrod, and Nir Ben-Tal. The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Research*, 37(suppl 1):D323–D327, January 2009.

[31] N. Goldman, J. L. Thorne, and D. T. Jones. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, 149:445–458, 1998.

[32] N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, 11:725–736,

1994.

[33] Nick Goldman, Jeffrey L. Thorne, and David T. Jones. Assessing the Impact of Secondary Structure and Solvent Accessibility on Protein Evolution. *Genetics*, 149(1):445–458, May 1998.

[34] B. J. Grant, A. P. C. Rodrigues, K. M. ElSawy, A. J. McCammon, and L. S. D. Caves. Bio3D: an R package for the comparative analysis of protein structures. *Bioinformatics*, 22:2695–2696, 2006.

[35] Geoffrey Grimmett and David Stirzaker. *Probability and Random Processes*. Oxford University Press, May 2001.

[36] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan. Protein sectors: Evolutionary units of three-dimensional structure. *Cell*, 138:774–786, 2009.

[37] B. Halle. Flexibility and packing in proteins. *Proc. Natl. Acad. Sci. USA*, 99:1274–1279, 2002.

[38] Bertil Halle. Flexibility and packing in proteins. *Proceedings of the National Academy of Sciences*, 99(3):1274–1279, February 2002.

[39] Thomas Hamelryck. An amino acid has two sides: A new 2d measure provides a different view of solvent exposure. *Proteins: Structure, Function, and Bioinformatics*, 59(1):38–48, April 2005.

[40] Tara Hessa, Hyun Kim, Karl Bihlmaier, Carolina Lundin, Jorrit Boekel, Helena Andersson, IngMarie Nilsson, Stephen H. White, and Gunnar von Heijne. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, 433(7024):377–381, January 2005.

[41] Konrad Hinsen. Physical arguments for distance-weighted interactions in elastic network models for proteins. *Proceedings of the National Academy of Sciences*, 106(45):E128–E128, November 2009.

[42] T.-T. Huang, M. L. del Valle Marcos, J.-K. Hwang, and J. Echave. A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. *BMC Evol. Biol.*, 14:78, 2014.

[43] Tsun-Tsao Huang, Mara L. del Valle Marcos, Jenn-Kang Hwang, and Julian Echave. A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. *BMC Evolutionary Biology*, 14(1):78, April 2014.

[44] Laurence D Hurst. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in Genetics*, 18(9):486–487, September 2002.

[45] E. L. Jackson, N. Ollikainen, A. W. Covert III, T. Kortemme, and C. O. Wilke. Amino-acid site variability among natural and designed proteins. *PeerJ*, 1:e211, 2013.

[46] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Mol. Biol. Evol.*, 31:736–749, 2014.

[47] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, July 1983.

[48] W Kabsch and C Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, December 1983. PMID: 6667333.

[49] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[50] M. Karplus and A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Struct. Biol.*, 9:646–652, 2002.

[51] K. Katoh, K.-I. Kuma, H. Toh, and T Miyata. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucl. Acids Res.*, 33:511–518, 2005.

[52] K. Katoh, K. Misawa, K.-I. Kuma, and T Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier

transform. *Nucl. Acids Res.*, 30:3059–3066, 2002.

[53] Kazutaka Katoh, Kei-ichi Kuma, Hiroyuki Toh, and Takashi Miyata. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, 33(2):511–518, January 2005.

[54] Motoo Kimura. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 267(5608):275–276, May 1977.

[55] S. L. Kosakovsky Pond, S. D. W. Frost, and S. V. Muse. HyPhy: hypothesis testing using phylogenetics. *Bioinformatics*, 21:676–679, 2005.

[56] S. Kryazhimskiy and J. B. Plotkin. The population genetics of dN/dS. *PLoS Genet.*, 4:e1000304, 2008.

[57] B. Kuhlman, G. Dantas, G.C. Ireton, V. Gabriele, and B.L. Stoddard. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302:1364–1368, 2003.

[58] Jack Kyte and Russell F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, May 1982.

[59] A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. Kaufman, D. P. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. Mandell, F. Richter, Y.-E. A. Ban, S. J.

Fleishman, J. E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popović, J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker, and P. Bradley. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology*, 487:545–574, 2011.

[60] Jie Liang and Ken A. Dill. Are Proteins Well-Packed? *Biophysical Journal*, 81(2):751–766, August 2001.

[61] H. Liao, W. Yeh, D. Chiang, R. L. Jernigan, and B. Lustig. Protein sequence entropy is closely related to packing density and hydrophobicity. *PEDS*, 18:59–64, 2005.

[62] H. Liao, W. Yeh, D. Chiang, R.L. Jernigan, and B. Lustig. Protein sequence entropy is closely related to packing density and hydrophobicity. *Protein engineering, design & selection : PEDS*, 18(2):59–64, February 2005.

[63] D. A. Liberles, S. A. Teichmann, I. Bahar, U. Bastolla, J. Bloom, E. BornbergBauer, L. J. Colwell, A. P. J. de Koning, N. V. Dokholyan, J. Echave, A. Elofsson, D. L. Gerloff, R. A. Goldstein, J. A. Grahnen, M. T. Holder, C. Lakner, N. Lartillot, S. C. Lovell, G. Naylor, T. Perica, D. D. Pollock, T. Pupko, L. Regan, A. Roger, N. Rubinstein, E. Shakhnovich, K. Sjölander, S. Sunyaev, A. I. Teufel, J. L. Thorne, J. W. Thornton, D. M. Weinreich, and S. Whelan. The interface of protein structure,

protein biophysics, and molecular evolution. *Protein Sci.*, 21:769–785, 2012.

[64] Chih-Peng Lin, Shao-Wei Huang, Yan-Long Lai, Shih-Chung Yen, Chien-Hua Shih, Chih-Hao Lu, Cuen-Chao Huang, and Jenn-Kang Hwang. Deriving protein dynamical properties from weighted protein contact number. *Proteins: Structure, Function, and Bioinformatics*, 72(3):929935, 2008.

[65] Y. Liu and I. Bahar. Sequence evolution correlates with structural dynamics. *Mol. Biol. Evol.*, 29:2253–2263, 2012.

[66] Ying Liu and Ivet Bahar. Sequence Evolution Correlates with Structural Dynamics. *Molecular Biology and Evolution*, 29(9):2253–2263, September 2012.

[67] B. Lorenz, I. Orgzall, and H.-O. Heuer. Universality and cluster structures in continuum models of percolation with two different radius distributions. *Journal of Physics A: Mathematical and General*, 26(18):4711, September 1993.

[68] S. Maguida, S. Fernandez-Albertia, and J. Echave. Evolutionary conservation of protein vibrational dynamics. *Gene*, 422:7–13, 2008.

[69] Mara Laura Marcos and Julian Echave. Too packed to change: site-specific substitution rates and side-chain packing in protein evolution. *bioRxiv*, page 013359, December 2014.

[70] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE*, 6:e28766, 2011.

[71] J. A. Marsh and S. A. Teichmann. Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *BioEssays*, 36:209–218, 2014.

[72] Itay Mayrose, Dan Graur, Nir Ben-Tal, and Tal Pupko. Comparison of Site-Specific Rate-Inference Methods for Protein Sequences: Empirical Bayesian Methods Are Superior. *Molecular Biology and Evolution*, 21(9):1781–1791, September 2004.

[73] A. G. Meyer, E. T. Dawson, and C. O. Wilke. Cross-species comparison of site-specific evolutionary-rate variation in influenza haemagglutinin. *Phil. Trans. R. Soc. B*, 368:20120334, 2013.

[74] A. G. Meyer and C. O. Wilke. Integrating sequence variation and protein structure to identify sites under selection. *Mol. Biol. Evol.*, 30:36–44, 2013.

[75] Austin G. Meyer, Eric T. Dawson, and Claus O. Wilke. Cross-species comparison of site-specific evolutionary-rate variation in influenza haemagglutinin. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 368(1614):20120334, March 2013.

[76] Austin G. Meyer and Claus O. Wilke. Integrating Sequence Variation and Protein Structure to Identify Sites under Selection. *Molecular Biology and Evolution*, 30(1):36–44, January 2013.

[77] Austin G. Meyer and Claus O. Wilke. Geometric constraints dominate the antigenic evolution of influenza H3n2 hemagglutinin. *bioRxiv*, page 014183, February 2015.

[78] L. A. Mirny and E. I. Shakhnovich. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.*, 291:177–196, 1999.

[79] Alexey G. Murzin, Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, April 1995.

[80] S. V. Muse and B. S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–724, September 1994.

[81] Z. Nevin Gerek, S. Kumar, and S. Banu Ozkan. Structural dynamics flexibility informs function and evolution at a proteome scale. *Evolutionary Applications*, 6:423–433, 2013.

[82] N. Ollikainen and T. Kortemme. Computational protein design quantifies structural constraints on amino acid covariation. *PLoS Comput. Biol.*, 9:e1003313, 2013.

[83] R. Orbach and M. Tachiki. Phonon-Induced Ion-Ion Coupling in Paramagnetic Salts. *Physical Review*, 158(2):524–529, June 1967.

[84] J. Overington, D. Donnelly, M. S. Johnson, A. Sali, and T. L. Blundell. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.*, 1:216–226, 1992.

[85] Giorgio Parisi and Nicolas Sourlas. Critical Behavior of Branched Polymers and the Lee-Yang Edge Singularity. *Physical Review Letters*, 46(14):871–874, April 1981.

[86] Kevin W Plaxco, Kim T Simons, and David Baker. Contact order, transition state placement and the refolding rates of single domain proteins1. *Journal of Molecular Biology*, 277(4):985–994, April 1998.

[87] Craig T. Porter, Gail J. Bartlett, and Janet M. Thornton. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research*, 32(suppl 1):D129–D133, January 2004.

[88] Tal Pupko, Rachel E. Bell, Itay Mayrose, Fabian Glaser, and Nir Ben-Tal. Rate4site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within

their homologues. *Bioinformatics (Oxford, England)*, 18 Suppl 1:S71–77, 2002.

[89] D. C. Ramsey, M. P. Scherrer, T. Zhou, and C. O. Wilke. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics*, 188:479–488, 2011.

[90] Duncan C Ramsey, Michael P Scherrer, Tong Zhou, and Claus O Wilke. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics*, 188(2):479–488, June 2011. PMID: 21467571.

[91] R. J. Read. Structure-factor probabilities for related structures. *Acta Crystallographica Section A Foundations of Crystallography*, 46(11):900–912, November 1990.

[92] Frederic M. Richards. The interpretation of protein structures: Total volume, group volume distributions and packing density. *Journal of Molecular Biology*, 82(1):1–14, January 1974.

[93] Michael A. Rodionov and Tom L. Blundell. Sequence and structure conservation in a protein core. *Proteins: Structure, Function, and Bioinformatics*, 33(3):358–366, November 1998.

[94] N. Rodrigue, N. Lartillot, D. Bryant, and H. Philippe. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene*, 347:207–217, 2005.

[95] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus. Hydrophobicity of amino acid residues in globular proteins. *Science*, 229(4716):834–838, August 1985. PMID: 4023714.

[96] D. Röthlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik, and D. Baker. Kemp elimination catalysts by computational enzyme design. *Nature*, 453:190–195, 2008.

[97] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comp. Phys.*, 23:327–341, 1977.

[98] Chris H. Rycroft. VORO++: A three-dimensional Voronoi cell library in C++. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 19(4):041111, December 2009.

[99] R. Salomon-Ferrer, A. W. Götz, D. Poole, S. Le Grand, and R. C. Walker. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.*, 9:3878–3888, 2013.

[100] Y. H. Sanejouand. Elastic network models: theoretical and empirical foundations. *Methods Mol. Biol.*, 924:601–616, 2013.

[101] M. P. Scherrer, A. G. Meyer, and C. O. Wilke. Modeling coding-sequence evolution within the context of residue solvent accessibility.

BMC Evol. Biol., 2012. submitted.

[102] Michael P. Scherrer, Austin G. Meyer, and Claus O. Wilke. Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evolutionary Biology*, 12(1):179, September 2012. PMID: 22967129.

[103] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The FoldX web server: an online force field. *Nucleic Acids Research*, 33(Web Server issue):W382–388, July 2005.

[104] Amir Shahmoradi, Dariya K. Sydykova, Stephanie J. Spielman, Eleisha L. Jackson, Eric T. Dawson, Austin G. Meyer, and Claus O. Wilke. Predicting Evolutionary Site Variability from Structure in Viral Proteins: Buriedness, Packing, Flexibility, and Design. *Journal of Molecular Evolution*, 79(3-4):130–142, September 2014.

[105] Peter S. Shenkin, Batu Erman, and Lucy D. Mastrandrea. Information-theoretical entropy as a measure of sequence variability. *Proteins: Structure, Function, and Bioinformatics*, 11(4):297–313, December 1991.

[106] C.-H. Shih, C.-M. Chang, Y.-S. Lin, W.C. Lo, and J.-K. Hwang. Evolutionary information hidden in a single protein structure. *Proteins: Structure, Function, and Bioinformatics*, 80:1647–1657, 2012.

[107] Chien-Hua Shih, Chih-Min Chang, Yeong-Shin Lin, Wei-Cheng Lo, and Jenn-Kang Hwang. Evolutionary information hidden in a single pro-

tein structure. *Proteins: Structure, Function, and Bioinformatics*, 80(6):1647–1657, June 2012.

[108] Tobias Sikosek and Hue Sun Chan. Biophysics of protein evolution and evolutionary protein biophysics. *Journal of The Royal Society Interface*, 11(100):20140419, November 2014.

[109] Noah Simon, Jerome Friedman, Trevor Hastie, Rob Tibshirani, et al. Regularization paths for coxs proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1–13, 2011.

[110] C. A. Smith and T. Kortemme. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J. Mol. Biol.*, 380:742–756, 2008.

[111] Alain Soyer, Jacques Chomilier, Jean-Paul Mornon, Rmi Jullien, and Jean-Franois Sadoc. Vorono\"\i Tessellation Reveals the Condensed Matter Character of Folded Proteins. *Physical Review Letters*, 85(16):3532–3535, October 2000.

[112] S. J. Spielman and C. O. Wilke. Membrane environment imposes unique selection pressures on transmembrane domains of G protein-coupled receptors. *J. Mol. Evol.*, 76:172–182, 2013.

[113] A. Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22:2688–2690, 2006.

[114] Dietrich Stauffer and Ammon Aharony. *Introduction To Percolation Theory*. CRC Press, July 1994.

[115] E. A. Stone and A. Sidow. Constructing a meaningful evolutionary average at the phylogenetic center of mass. *BMC Bioinformatics*, 8:222, 2007.

[116] Y. Suzuki. Natural selection on the influenza virus genome. *Mol. Biol. Evol.*, 23:1902–1911, 2006.

[117] M. Z. Tien, A. G. Meyer, D. K. Sydykova, S. J. Spielman, and C. O. Wilke. Maximum allowed solvent accessibilites of residues in proteins. *PLOS ONE*, 8:e80635, 2013.

[118] Matthew Z. Tien, Austin G. Meyer, Dariya K. Sydykova, Stephanie J. Spielman, and Claus O. Wilke. Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLoS ONE*, 8(11):e80635, November 2013.

[119] Neil R. Voss and Mark Gerstein. 3v: cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Research*, 38(Web Server issue):W555–W562, July 2010.

[120] Ivar Waller. Zur Frage der Einwirkung der Wrmebewegung auf die Interferenz von Rntgenstrahlen. *Zeitschrift fr Physik*, 17(1):398–408, December 1923.

[121] E. C. Webb. Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry

and Molecular Biology on the Nomenclature and Classification of Enzymes. (Ed. 6):xiii + 863 pp., 1992.

[122] C. O. Wilke and D. A. Drummond. Signatures of protein biophysics in coding sequence evolution. *Cur. Opin. Struct. Biol.*, 20:385–389, 2010.

[123] Claus O. Wilke, Jesse D. Bloom, D. Allan Drummond, and Alpan Raval. Predicting the Tolerance of Proteins to Random Amino Acid Substitution. *Biophysical Journal*, 89(6):3714–3720, December 2005.

[124] W. C. Wimley and S. H. White. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nature Structural Biology*, 3(10):842–848, October 1996.

[125] Fei Xia, Dudu Tong, Lifeng Yang, Dayong Wang, Steven C. H. Hoi, Patrice Koehl, and Lanyuan Lu. Identifying essential pairwise interactions in elastic network model using the alpha shape theory. *Journal of Computational Chemistry*, 35(15):1111–1121, May 2014.

[126] Lei Yang, Guang Song, and Robert L. Jernigan. Protein elastic network models and the ranges of cooperativity. *Proceedings of the National Academy of Sciences*, 106(30):12347–12352, July 2009.

[127] Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10(6):1396–1401, November 1993.

[128] Z. Yang. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J. Mol. Evol.*, 51:423–432, 2000.

[129] Ziheng Yang. *Computational Molecular Evolution*. OUP Oxford, October 2006.

[130] S.-W. Yeh, T.-T. Huang, J.-W. Liu, S.-H. Yu, C.-H. Shih, J.-K. Hwang, and J. Echave. Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. *BioMed Research International*, 2014:572409, 2014.

[131] S.-W. Yeh, J.-W. Liu, S.-H. Yu, C.-H. Shih, J.-K. Hwang, and J. Echave. Site-specific structural constraints on protein sequence evolutionary divergence: Local packing density versus solvent exposure. *Mol. Biol. Evol.*, 31:135–139, 2014.

[132] So-Wei Yeh, Tsun-Tsao Huang, Jen-Wei Liu, Sung-Huan Yu, Chien-Hua Shih, Jenn-Kang Hwang, and Julian Echave. Local Packing Density Is the Main Structural Determinant of the Rate of Protein Sequence Evolution at Site Level. *BioMed Research International*, 2014:e572409, July 2014.

[133] So-Wei Yeh, Jen-Wei Liu, Sung-Huan Yu, Chien-Hua Shih, Jenn-Kang Hwang, and Julian Echave. Site-Specific Structural Constraints on Protein Sequence Evolutionary Divergence: Local Packing Density versus

Solvent Exposure. *Molecular Biology and Evolution*, 31(1):135–139, January 2014.

[134] W. Zhou and H. Yan. Alpha shape and Delaunay triangulation in studies of protein-related interactions. *Briefings in Bioinformatics*, 15(1):54–64, January 2014.

[135] Afra Zomorodian, Leonidas Guibas, and Patrice Koehl. Geometric filtering of pairwise atomic interactions applied to the design of efficient statistical potentials. *Computer Aided Geometric Design*, 23(6):531–544, August 2006.

Vita

Amir Shahmoradi attended Mofid Educational Institute and the National Organization for Development of Exceptional Talents in Tehran, Iran where he received his diploma in Mathematics & Physics. He then attended Sharif University of Technology in Tehran where he obtained a Bachelor of Science in Physics in August of 2007 with special focus in High Energy Astrophysics and Particle Physics under the directions of Professor Jalal Samimi and Professor Mahmoud Bahmanabadi. Upon graduation, he joined the Physics program at Michigan Tech. University where he completed a Master of Science in Physics in May of 2011 under the directions of Professor Robert Nemiroff. In June 2011, he began his graduate studies in the field of Plasma sciences at the national Institute for Fusion Studies (IFS) at the University of Texas at Austin (UT Austin) and later in the field of Biophysics under the joint directions of Professor Swadesh Mahajan at IFS and Professor Claus Wilke at Institute for Cellular and Molecular Biology (ICMB) at UT Austin. In addition to his PhD candidacy in Physics, he started his Doctoral Portfolio Degree in Computational Sciences – a joint program of Texas Advanced Computing Center (TACC) and the Department of Statistics and Data Science at UT Austin – in Fall 2013. After completing his Ph.D., he expects to expand his understanding of natural phenomena by continuing quantitative research in diverse fields of science, in particular Physics, Biology and Astronomy.

Permanent address: a.shahmoradi@gmail.com;
amir@physics.utexas.edu

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.